

# 國立交通大學

資訊工程學系

## 碩士論文

以小波轉換為基礎的可調式視壓縮之  
流量控制機制



A Model-based Rate Allocation Mechanism for  
Wavelet-based Embedded Coding

研究生：游雅惠

指導教授：蔡淳仁 博士

中華民國九十四年六月

以小波轉換為基礎的可調式視壓縮之流量控制機制

**A Model-based Rate Allocation Mechanism for  
Wavelet-based Embedded Coding**

研究生：游雅惠

Student : Ya-Hui Yu

指導教授：蔡淳仁

Advisor : Chun-Jen Tsai

國立交通大學

資訊工程系



Submitted to Department of Computer Science and Information Engineering

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of Master

in

Computer Science and Information Engineering

June 2005

Hsinchu, Taiwan, Republic of China

中華民國九十四年六月

# 以小波轉換為基礎的可調式視壓縮之 流量控制機制

## 摘要

在以小波為基礎的嵌入式的圖像或影像壓縮過程中，可調變的特性和壓縮效能之間的平衡點可由階層式的封裝，即對於不同階層給予特定的壓縮比的方式達到部份的最佳化。這個過程稱為流量控制機制。一般典型的流量控制機制可歸類為兩個重要的過程，一是壓縮大小和失真度之間的最佳化問題，另一為簡單但收斂速度慢的搜尋方法。而在此篇論文中所提出的提案目標，為一個以建立模組化的方式達成高度效能的流量控制系統。此演算法的基本精神為以動態分析的過程，精確地建立圖像或影像的壓縮大小和失真度之間的關係。實驗為在一套嵌入式可調變的壓縮、解壓縮的方法中實作此演算法，結果顯示搜尋所須的時間明顯減少，且所取得的壓縮內容仍接近最佳化的效能。此技術可應用至各類以小波為基礎的嵌入式圖像或影像壓縮過程，如 JPEG2000 及以 MCTF 架構為基礎的可調變式影像壓縮方法。

# **A Model-based Rate Allocation Mechanism for Wavelet-based Embedded Image and Video Coding**

## ***Abstract***

In wavelet-based embedded coding for still images and/or videos, a trade-off between scalability and coding efficiency is achieved via layered-packetization of the embedded bitstreams optimized for several target operating bitrate points. This process is called rate allocation (tier-2 coding). The typical rate allocation mechanism is formulated as a rate-distortion optimization problem and a simple searching method with slow convergence rate is used. In this paper, a highly efficient model-based rate allocation mechanism is proposed. The algorithm is based on adaptive analysis of the relationship between source-coding rate and distortion of the image/video data. Experiments conducted on a scalable video codec show that the proposed technique greatly reduces the search time for the optimal solution. The techniques can be applied to various wavelet-based embedded image and video coding schemes, such as the JPEG2000 codec and the MCTF-based scalable video codecs.

## Acknowledgement

I am glad to finalize this thesis by the support of many people. First, my advisor, Chun-Jen Tsai, gives me a lot of motivation and suggestions. Also, he encourages me to think in various viewpoints to create new ideas. Then, I appreciate the great help and comments from my seniors, juniors, and classmates. Of course, I would like to thank my family for giving me strong economic and moral support. Finally, it is my pleasure to finish this thesis in my laboratory, MMES Lab of CSIE in NCTU.



# Content

<b>1. Introduction .....</b>	<b>10</b>
<b>2. Previous Work.....</b>	<b>13</b>
2.1. Rate Distortion Characteristics Model.....	14
2.1.1. Discrete R-D Model.....	14
2.1.2. Parameterized Closed-form Model.....	15
2.2. Bit Allocation .....	15
2.2.1. Lagrange Multiplier Optimization.....	16
2.2.2. Iterative search method .....	17
2.2.3. Fast search method using special data structure.....	17
2.3. Shortcomings of Existing Work .....	18
<b>3. Problem Formulation and Analysis.....</b>	<b>20</b>
3.1. Introduction to Wavelet-based Embedded Coding .....	20
3.1.1. Overall Scheme of Wavelet-based Embedded Coding .....	21
3.1.2. General Rate Control Algorithm .....	22
3.2. Basis of the Proposed Formulation.....	23
3.2.1. Rate Distortion Function.....	23
3.2.2. Lagrange Multiplier Optimization.....	25
3.3. Analysis of Source Coding Information Theory .....	27
3.3.1. Rate Distortion Function .....	27
3.3.2. Lagrange Multiplier Optimization.....	29
3.4. Implementation Issues .....	32
<b>4. Proposed Rate Control Framework .....</b>	<b>34</b>
4.1. General Rate Control Extractor.....	34
4.2. Proposed Rate Control Extractor.....	37
4.2.1. R-Lambda Model Analysis .....	38
4.2.2. Overall Framework.....	42
4.2.3. Lambda Search Procedure.....	45
4.3. Proposed Multiple Adaptation Scheme.....	47
<b>5. Experimental Results.....</b>	<b>49</b>
5.1. Computational Cost Reduction in Rate Control Extractor .....	49

5.2. Side Information Saving for Multiple Adaptation Scheme.....55

**6. Conclusion and Future Work ..... 56**

**7. References ..... 60**



## List of Figures

Fig 1. Multiple Adaptation Scheme .....	11
Fig 2. General wavelet coding framework. ....	21
Fig 3. General algorithm for quality layer formation.....	22
Fig 4. The Lagrangian cost function for the operating points with different rate and distortion values .....	27
Fig 5. R-D Model for Coding Blocks in Codec MSRA.....	29
Fig 6. Frame 254 in the video sequence Stefan .....	31
Fig 7. The RD curves corresponding to two MBs in Fig 6. ....	31
Fig 8. The enlarged RD curve around the optimal mode in Fig 7 for the MB on Stefan's foot.....	32
Fig 9. The enlarged RD curve around the optimal mode in Fig 7 for the MB on the tennis court.....	32
Fig 10. Flow Chart of MSRA Rate Control Extractor .....	34
Fig 11. Bitstream Format.....	35
Fig 12. Overall Computation Cost Analysis.....	36
Fig 13. Truncation for Resolution Scalability.....	37
Fig 14. Truncation for Frame Rate Scalability.....	37
Fig 15. Truncation for Bitrate (Quality) Scalability.....	37
Fig 16. The Concept of the Proposed Rate Control Extractor.....	38
Fig 17. Coding Block level R-Lambda Relationship Examination .....	40
Fig 18. GOP level R-Lambda Relationship Examination .....	41
Fig 19. Overall Framework of the Proposed Rate Control Mechanism.....	44
Fig 20. Search Path for Bisection Method.....	46
Fig 21. Search Path for the Proposed Method .....	46
Fig 22. Multiple Adaptation Behavior.....	47
Fig 23. Saving Ratio of the Iteration Times.....	50
Fig 24. Overall Computational Cost Saving for Rate Control Extractor .....	50
Fig 25. PSNR Performance Comparison of Stefan.....	51
Fig 26. PSNR Performance Comparison of Football.....	52
Fig 27. PSNR Performance Comparison of Foreman .....	53
Fig 28. Frame Level PSNR Performance Comparison of Foreman .....	54
Fig 29. Bit Allocation of the Second GOP in Foreman with MSRA Codec.....	54
Fig 30. Bit Allocation of the Second GOP in Foreman with the Proposed Method ....	54
Fig 31. Bit Allocation Comparison of the Second GOP in Foreman.....	55



## List of Tables

Table 1.	Number of Iterations Comparison for Lambda Search.....	49
Table 2.	Side Information Saving Ratio .....	55
Table 3.	Algorithm Comparison.....	57
Table 4.	Rate Control Mechanism Comparison.....	58

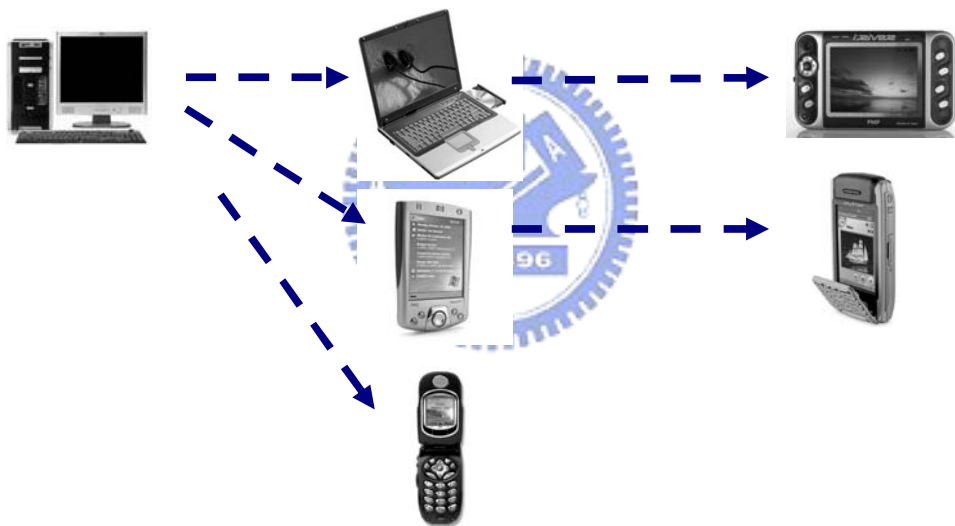


# 1. Introduction

Data networks for multimedia communications are growing fast nowadays. The environment varies from broadband cable/ADSL networks, dial-up connections, to wireless/mobile networks. Today, data transmission over Internet is common in the forms of pure text, audio or video. In the future, multimedia information will be more popular for entertainment, education or business purposes. The expectation of visual quality of the content will be higher than what we have over Internet today. In addition, the terminal devices of the transmission system are diversified as well. Also, media storage ranges from low capacity flash-based memory card to high capacity DVD. Finally, the display monitor of the device may be a small size screen on a mobile device or a high definition projection system.

For different applications on various devices or under different network conditions, the available bandwidth and resource may be highly divergent. Therefore, there are many multimedia transmission techniques proposed to overcome different application scenarios. In order to adapt to the dynamic network environment, an error control module is used to conceal the error resulted from lossy transmission while a rate adaptation module is used to adapt the bitrate of the multimedia data to the network bandwidth. For efficient real-time adaptation, the rate adaptation module relies on the design of an embedded scalable codec to achieve best runtime quality. A common approach for scalable bitstream is to use a layered coding approach such as MPEG Simple Scalable Profile or FGS. The content can achieve best quality at certain bitrates conditions. When the adaptation must take into account device properties (such as screen size) and the quality requirement is higher, a fully embedded bitstream is a better way to achieve the goal.

The most desirable property of embedded bitstream is that the video parameters such as resolution, frame rate, and quality of the bitstream can be dynamically selected after the encoding procedures. And one of the nice features of such embedded bitstreams is called multiple-adaptation. For example (Fig 1), the workstation can transmit different scalable bitstream to different devices. Upon reception of the embedded bitstreams, the notebook (or the PDA) can truncate the bitstream further without decoding procedures and send it to another device with a smaller screen. Hence, the technique is suitable for the communication in the various transmission environments including the bandwidth of the internet and the properties of the receiving devices.

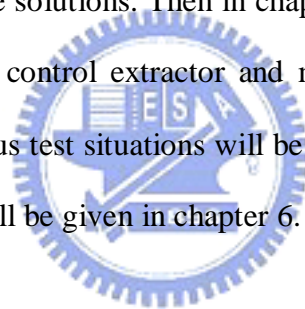


**Fig 1. Multiple Adaptation Scheme**

The operations of the truncation can be implemented by two approaches. The first one is that the workstation provides a layer structured bitstream, and the notebook extracts the embedded layers which do not exceed the transport and terminal constraints of the target device. This approach is quite simple but the bitstream can not achieve the best quality possible. And the second approach is that the desktop provides a fully embedded bitstream, the notebook can extract the bitstream according to the constraint of the target device. This approach is more precise than the previous

one, but the side information, namely the RD information, is required and the complexity of the extractor is much higher. The issue is especially true for resource critical systems, like the handheld device, PDA or cellular phone. Therefore, a rate allocation mechanism which can extract embedded bitstream with accurate target bit-rate in tolerable time and within distinct memory usage is very important.

As a result, the goal of this thesis is to design a rate control extractor for multiple adaptation applications which needs less side information and also reduces the computational cost. The organization of the thesis is as follows. Chapter 2 introduces some previous work of the rate control scheme for embedded codec works and discusses their strength and weakness. Chapter 3 formulates the problem and the fundamental theory behind the solutions. Then in chapter 4, the proposed method will be elaborated, including rate control extractor and multiple adaptation design. The experimental results for various test situations will be shown in chapter 5. Finally, the conclusion and discussions will be given in chapter 6.



## 2. Previous Work

An embedded image/video codec is different from the traditional image/video codec in the way that an embedded codec generates compressed bitstreams that can be further reduced to a smaller target bitstream for different application scenarios. The first well-known wavelet-based embedded image coder is the embedded zero-tree wavelet compression (EZW) technique, proposed by Shapiro [1]. Later, the basic concept is then extended to other embedded coding schemes. For example, Taubman and Zakhor proposed a Layered Zero Coding (LZC) method [3] in 1994, Said and Pearlman proposed an algorithm that performs Spatial Partitioning of Images into Hierarchical Trees (SPIHT) [2] in 1996, and Taubman proposed Embedded Block Coding with Optimized Truncation (EBCOT) [4] in 2000. EBCOT is adopted by the well-known image coding standard, JPEG2000.

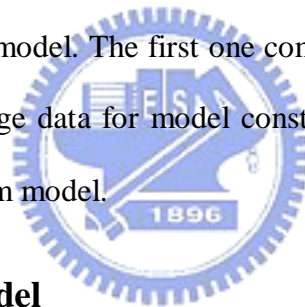
For embedded image and/or video coding, rate control is a crucial module after entropy coding. The purpose of rate control is to extract a smaller sub-bitstream from a compressed bitstream that meets some application criteria. During the rate control process, the frame rate, resolution, and bitrate can all be changed to form the target bitstreams. This is called full-dimensional scalability. A rate control algorithm plays an important role to achieve full scalability in an optimal way.

Many rate control schemes have been proposed for embedded image/video codecs. The basic idea behind these rate control techniques is similar. In general, the rate control scheme for embedded coders is composed of two parts. The first part is to model the rate-distortion characteristics of a group of input image/video data, and the second part is the bit allocation mechanism that assigns proper number of bits to various parts of the input data according to their importance. Since similar techniques

can be applied to both image and video codecs, the remaining discussions in this chapter do not distinguish between these two types of applications.

## 2.1. Rate Distortion Characteristics Model

For wavelet-based codecs, image/video data is partitioned into coding units, which could be a frame, a frequency band, or a coding block. Several rate-distortion (R-D) modeling techniques have been proposed to establish the R-D relationship for each coding unit [5]-[9]. An R-D model represents the degree of degradation of a coding unit when the size of the compressed data is constrained by the available bandwidth. The R-D models of the coding units can be used by the bit allocation algorithm to sort out the priority of the coding units. There are two typical ways to build the R-D characteristics model. The first one computes discrete R-D relationship data points from the real image data for model construction. The other method is to use a parameterized close-form model.



### 2.1.1. Discrete R-D Model

In wavelet-based embedded codecs, bitrate scalability is achieved by fractional bit plane coding. As the available bandwidth of the target applications goes from low to high, more and more fractional bit planes could be included into the target bitstreams according to their significance. In other words, an embedded bitstream is composed of fractional bit planes, and the more fractional bit planes the bitstream contains, the higher quality it would be. Therefore, inclusion of an additional fractional bit plane in a coding unit contributes to both the increase of bits (rate) and reduction of quality loss (distortion). Recording of the rate and distortion data point of each fractional bit plane provides a precise yet discrete R-D model of the embedded bitstream [8] [9].

By using real data points to represent the R-D model of the image/video data, the tradeoff between rate and distortion at each truncation point can be precisely determined. However, storing all the discrete R-D values for each fractional bit plane in each coding unit during the bit allocation procedure requires a lot of memory space, especially for video coding. Furthermore, in order to find the best truncation point which matches the rate constraint, some search techniques (possibly time-consuming) must be applied while doing bit allocation.

### **2.1.2. Parameterized Closed-form Model**

Different from the discrete R-D model approach, some literatures [5]-[7] use close-form models to describe the R-D characteristic of the image/video data. This approach first applies information theory on a simplified source model and a codec model to calculate the relationship between coding rate and distortion. In the closed form R-D equation, content-dependent information is summarized in a few parameters. With the parameterized R-D model, the R-D characteristic of each coding unit will be estimated at runtime by solving the content-dependent parameters. In general, the parameters can be estimated from the content statistics and/or by curve fitting of sparse data points.

By using a closed-form R-D model, memory consumption of the rate control process can be substantially reduced, but the accuracy of bit allocation may decrease, depending on the accuracy of the R-D model.

## **2.2. Bit Allocation**

The goal of the bit allocation procedure is to achieve maximal quality for a given bitrate or minimal bitrate for a given distortion. Given the R-D characteristics models for each coding unit, nonlinear optimization techniques can then be applied to

distribute the coding bits among all coding unit in an optimal way. A popular approach is to use the Lagrange multiplier to transform constrained optimization problem into unconstrained optimization problem. During this process, some truncation points will be deleted from the candidates of optimal solutions since they do not falls on the convex hull of R-D curves. The problem of bit allocation now comes down to determine which combinations of all possible optimal truncation points can meet the target bitrate best. The procedure is usually conducted in one of two ways. The first one is to use an iterative search method to find the best combination of the bitrate in all coding blocks by trial and error. The other approach is to design special data structure to store extra information during Lagrange multiplier optimization for quick location of the optimal truncation points for bit allocation.

### **2.2.1. Lagrange Multiplier Optimization**

For each coding unit, the Lagrange multiplier optimization method is used to achieve better rate distortion tradeoff [5]-[9]. In this approach, a cost function of a constrained optimization problem (of rate vs. distortion) is converted to an unconstrained optimization problem using the Lagrange multiplier formulation. The optimal solutions (for given constraints) of this cost function are located on the convex hull of the rate-distortion curve. In another word, the tangent values of the R-D curve at all the optimal solution points should get smaller as the bitrate increases (assuming that the R-D function has bitrate as the domain axis). The truncated points which do not follow the rule are not valid optimal truncation candidates.

With this optimization rule, the non-optimal truncated points at coding unit level can be eliminated. The frame level bit allocation module for image coder or the GOP level bit allocation module for video coder can then focus on selecting only optimal truncation points from all the coding units in order to meet the target bitrate or target



distortion constraints.

### **2.2.2. Iterative search method**

After the establishment of the R-D characteristic model and the optimization process using Lagrange multiplier, each optimal truncation point contains three attributes including rate, distortion, and the Lagrange multiplier value (refer to as the  $\lambda$  value hereafter). The next step is to form an optimal target bitstream given a rate or distortion constraint. Some literatures use iterative search method to achieve this goal [6]-[9]. Among the optimal truncation point attributes, the  $\lambda$  values represent the trade off parameters between rate and distortion at those truncation points. By applying a specific  $\lambda_c$  to all coding units, the collective set of all truncation points with their  $\lambda$  values closest to  $\lambda_c$  builds an optimal bitstream with the given constraint. Simply put, given  $\lambda_c$ , an optimal bitstream fulfilling some rate (or distortion) constraint can be generated (which may not exactly match the target constraint). An iterative search method, such as bisection search, can be used to iteratively selecting different  $\lambda_c$  until the composed bitstream meets the target constraint.

The iterative search method can create a bitstream with its bitrate or distortion close to the target constraint. The weakness of the iterative search method is that the convergence rate may be slow. Further improvement can be achieved if the search process takes advantage of the R-D characteristics of the content.

### **2.2.3. Fast search method using special data structure**

Besides the iterative search method, some studies ([5], [10]) designed special data structure to record R-D tradeoff points of all coding units. For example, a heap-based structure has been proposed to process rate allocation for embedded image coding [10]. The heap structure which contains all possible truncation points is built

internally during encoding process and some heap manipulations, such as shift-down and update root, are conducted according to rate distortion property of each truncation point. The heap manipulation operations stop when the heap tree is balanced and the root of the tree meets the target bitrate constraint. At this point, the final bitstream is composed. Another approach uses quadtree merge-based algorithm is proposed in [5]. Similar to the heap-based proposal, this method tries to achieve fast R-D optimization by applying simple operations to manipulate the data structure during the bit allocation process.

One major disadvantage of fast search algorithm with well-designed data structure is that the memory required may be extremely large in order to build the complete data structure to store all coding units information, especially for video coding.

### **2.3. Shortcomings of Existing Work**

For the two stages of embedded coding rate control algorithms, there is still plenty of room for improvements, either in model accuracy or in computational complexity reduction. In the first phase of rate control, namely building R-D characteristics model, using discrete R-D relationship data points can represent the real rate distortion data well, but the memory requirement is pretty high. For the methods of using closed form models, the precision of hitting the target constraint (rate or distortion) depends on the accuracy of the model. Therefore, in order to meet the target constraint with high precision and low complexity, an accurate closed form model based on more elaborated theoretical analysis is necessary.

In the second phase of rate control, namely the bit allocation procedure, neither iterative search method nor fast data structure approach takes advantage of the characteristics of the content. Although these types of bit allocation are accurate, the

computational complexity is high. A content-adaptive bit allocation scheme should be developed in order to reduce the computational complexity while maintain the accuracy of the solution.



### **3. Problem Formulation and Analysis**

The goal of this thesis is to formulate an embedded codec rate control algorithm which can well explore full-dimensional scalability, namely spatial (resolution) scalability, temporal (frame rate) scalability and SNR (quality) scalability. The proposed mechanism is designed for low-complexity systems and achieves low memory usage with high computational efficiency while maintaining same quality. In addition, the proposed scheme is more suitable for multiple adaptation applications.

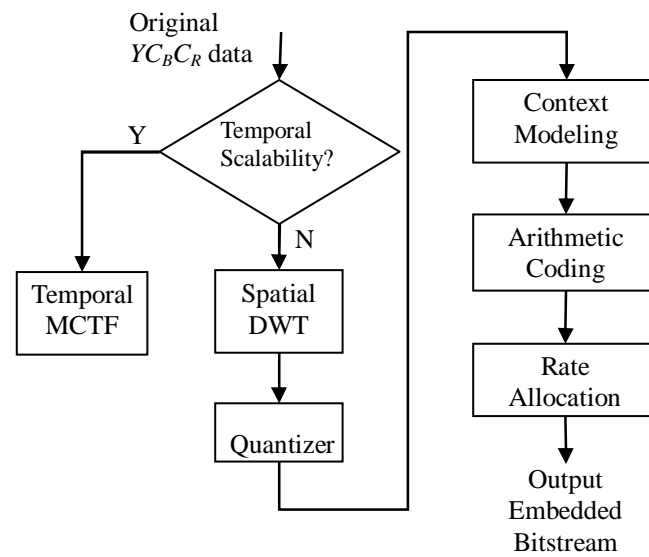
To design a scalable rate control algorithm, first, we must derive an R-D model which can adaptively and accurately describe the characteristics of image/video contents. The derivation of the model is based on information theory. Secondly, it is crucial to adopt a proper optimization technique that works well with the model. Since a well-designed bit allocation procedure should achieve better trade-off between rate and distortion efficiently, the optimization technique becomes an essential element to accomplish the goal.

In the following sections, the common embedded coding scheme is first introduced. The concepts of related information theories are then explained, and the conditions of using the theories in source coding are also analyzed. Finally, the critical parts for solving the problem are discussed.

#### **3.1. Introduction to Wavelet-based Embedded Coding**

In this section, the coding procedure of a general wavelet-based embedded codec is introduced. The components related to the proposed rate control mechanism are presented, and the input and output data format is defined. In addition, the architecture of a general rate control mechanism is illustrated as a basis of the proposed algorithm.

### 3.1.1. Overall Scheme of Wavelet-based Embedded Coding



**Fig 2. General wavelet coding framework.**

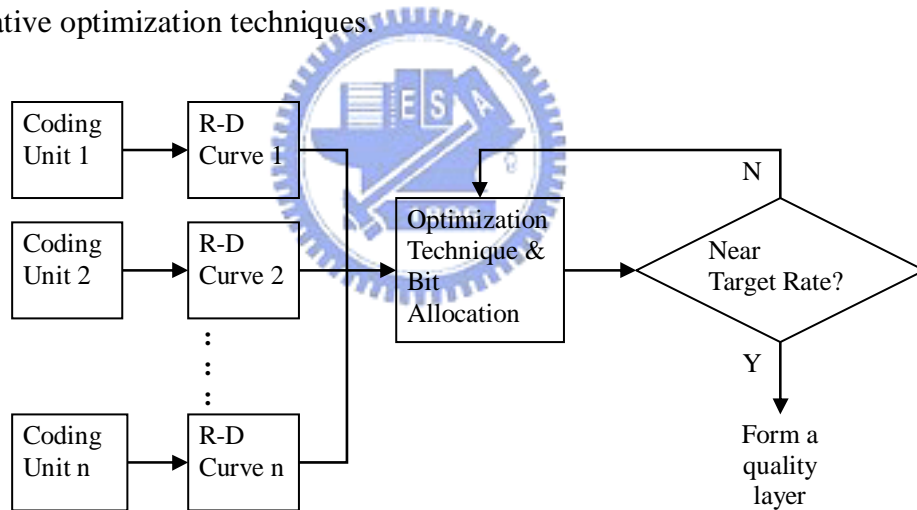
A general framework for wavelet-based embedded image/video coding [4] [8] [9] is shown in Fig 2. The input  $YCbCr$  frame data is first transformed into frequency domain via temporal (for video coding only) and spatial (for both image and video coding) subband decomposition. The transform process is followed by the quantization process and the entropy coder with rate allocation mechanism. Popular wavelet-based image and video coders typically use Discrete Wavelet Transform (DWT) for spatial subband decomposition and Motion-Compensated Temporal Filtering (MCTF) for temporal subband decomposition. Arithmetic coding with adaptive context modeling is adopted as the entropy coder. And the rate allocation procedure is used to explore bitrate (quality) scalability of the embedded bitstreams. The input to the rate allocation mechanism is the complete entropy-coded embedded bitstream while the output is its subset which matches the target criterions.

The differences between embedded image coding and video coding are mainly in the application of MCTF in temporal axis and the context models for each subband in the entropy coder. The rate allocation mechanism, the main topic of this paper, can be

applied to both wavelet image and video codecs in a similar manner.

### 3.1.2. General Rate Control Algorithm

After temporal and spatial wavelet transform, each decomposed frequency subbands are coded independently, and the pixels in a subband are split into coding units. The R-D curve which characterizes the content of a coding unit is established by achievable rate distortion points and is formed during entropy coding procedure. The rate allocation process then tries to find the optimal embedded points on the R-D curve according to the prospective target bitrates. The formation of a quality layer with a given bitrate criterion is illustrated in Fig 3. The optimal truncated bitstream that matches the given target bitrate with minimum distortion is typically found by using iterative optimization techniques.



**Fig 3. General algorithm for quality layer formation**

The proposed rate control scheme follows the general wavelet coding framework. Nevertheless, the proposed algorithm reduces memory usage by deriving an efficient R-D model and slashes computational cost by adopting a better bit allocation mechanism.

## 3.2. Basis of the Proposed Formulation

In this section, the fundamental theories which will be used for the development of the proposed algorithm are introduced. There are two procedures in a generic wavelet-based scalable rate control mechanism, namely, the construction of the rate distortion model and the bit allocation process. The first procedure is based on rate distortion function analysis, while the second procedure usually utilizes the Lagrange multiplier optimization technique.

### 3.2.1. Rate Distortion Function

The concept of rate distortion function is first published by Shannon in 1948 in his famous paper on Information Theory [11]. The problem is discussed based on a classical communication system, which consists of source, source coder, channel coder, channel, channel decoder, source decoder and destination. The source content is transmitted from source to destination through the coders and the channel. The original source bits is represented as  $\mathbf{b}$ , and the content bits received by the destination side is symbolized as  $\mathbf{b}'$ . The distortion made from the source coder, channel coder and the channel noise is represented as  $\mathbf{d}(\mathbf{b}, \mathbf{b}')$  which is measured by certain distortion criterion. Because the distinct rate distortion function is difficult to compute, the approximation can usually be described by various bounds. The best known of these bounds, the “Shannon Lower Bound”, is designed for continuous amplitude  $\mathbf{b}$  and  $\mathbf{b}'$  and is given as follows. The  $E(x)$  function represents the entropy of the signal  $x$ .

$$R(D) \geq E(\mathbf{b}) - E(\mathbf{d}(\mathbf{b}, \mathbf{b}')),$$
$$E(x) = \int p(x) \log_2 p(x) dx$$
(1)

The equation gives the lower bound of the channel capacity required to obtain the content with distortion lower than  $d(\mathbf{b}, \mathbf{b}')$  on the destination side.

By using the basic form, several literatures elaborate the equation with different distortion measure criterion and various probability density functions for the source [12] [13] [14]. Eq (2) shows the function with the distortion criterion, square error, which is a frequently used measurement in source coding.

$$R(D) \geq E(b) - \frac{1}{2} \log_2(2\pi e D) \quad (2)$$

The following equations give a general form of the square-error-criterion rate distortion function by the inference of Eq (2).

$$\begin{aligned}
 R(D) &\geq E(b) - \frac{1}{2} \log_2(2\pi e D) = R_L(D) \\
 R_L(D) &= \frac{1}{2} \log_2 2^{2E(b)} - \frac{1}{2} \log_2(2\pi e D) \\
 &= \frac{1}{2} \log_2 \frac{2^{2E(b)}}{2\pi e D}, \quad \omega = \frac{2^{2E(b)}}{2\pi e} \\
 &= \frac{1}{2} \log_2 \frac{\omega}{D}
 \end{aligned} \quad (3)$$

The parameter,  $\omega$ , in Eq (3) is related to the probability density function of the source signal. Take Gaussian distribution for example, assume the arbitrary mean is  $\mu$ , the variance is  $\sigma^2$ , and the probability density function is

$$p(x) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \quad (4)$$

The entropy of the source signal with Gaussian distribution is



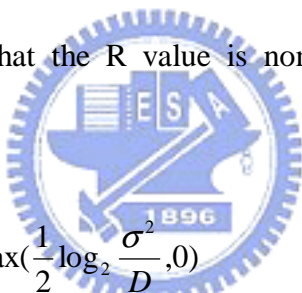
$$E(p) = \frac{1}{2} \log_2 2\pi e \sigma^2. \quad (5)$$

Therefore,

$$\begin{aligned} \omega &= \frac{2^{2E(b)}}{2\pi e} \\ &= \frac{2\pi e \sigma^2}{2\pi e}, \text{ and} \\ &= \sigma^2 \end{aligned} \quad (6)$$

$$\begin{aligned} R_L(D) &= \frac{1}{2} \log_2 \frac{\omega}{D} \\ &= \frac{1}{2} \log_2 \frac{\sigma^2}{D}. \end{aligned}$$

In order to make sure that the R value is nonnegative, the equation can be rewritten as



$$\begin{aligned} R_L(D) &= \max\left(\frac{1}{2} \log_2 \frac{\sigma^2}{D}, 0\right) \\ &= \begin{cases} \frac{1}{2} \log_2 \frac{\sigma^2}{D}, & 0 \leq D \leq \sigma^2 \\ 0, & D \geq \sigma^2 \end{cases} \end{aligned} \quad (7)$$

In this section, the general rate distortion relationship is established. The rate distortion model can be extended by using different distortion measures or content probability density functions.

### 3.2.2. Lagrange Multiplier Optimization

In the bit allocation procedure, the goal is to generate a final bitstream which matches the criterion of the target bitrate and also minimizes the distortion. The general format of the problem can be described below as a budget-constrained allocation problem [15].

Find the optimal quantizer, or operating point,  $x(i)$  for each coding unit  $i$  such that

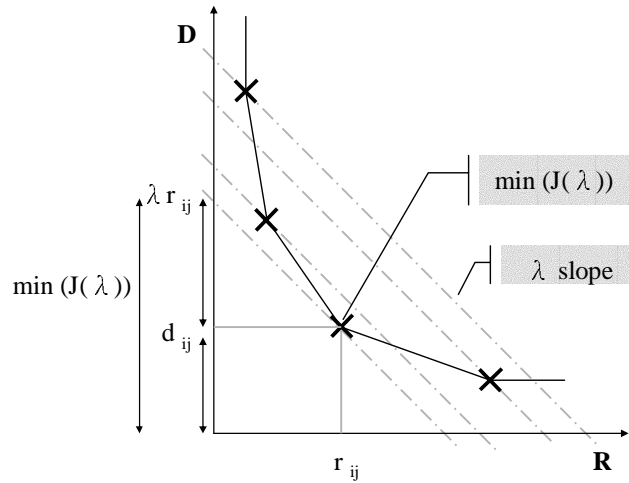
$$\sum_{i=1}^N r_{ix(i)} \leq R_T \quad (8)$$

and some metric  $f(d_{1x(1)}, d_{2x(2)}, d_{3x(3)}, \dots, d_{Nx(N)})$  is minimized.

The classical solution for the problem is based on the Lagrange optimization technique first introduced by Everett [16]. This approach is used for a source coding application in the beginning and applied to various constrained allocation problem later. The basic idea is described as follows. For a given Lagrange multiplier  $\lambda$  which is nonnegative, the Lagrangian cost function for each coding unit  $i$  under certain operating criterion  $j$  is formulated in Eq (9).

$$J_{ij}(R) = d_{ij} + \lambda r_{ij} \quad (9)$$

The concept of the Lagrangian cost is illustrated in Fig 4. When the  $\lambda$  value is close to zero, the tangent line (with slope  $\lambda$ ) of the optimal solution is almost like a horizontal line. As a result, minimizing Lagrangian cost,  $J$ , is equal to minimizing the distortion. On the other hand, minimizing  $J$  function when  $\lambda$  becomes rather large is equivalent to minimizing the rate. In another words,  $\lambda$  value plays a role to define the trade-off between rate and distortion. For a given  $\lambda$ , the operating point which has the nearest trade-off property will be selected by minimizing the Lagrangian cost function.



**Fig 4. The Lagrangian cost function for the operating points with different rate and distortion values**

By adopting Lagrange multiplier optimization technique, Eq (8) is simply transformed into an unconstrained problem as Eq (9). But the constrained budget should still be taken care of by defining a proper  $\lambda$  value. Consequently, the determination of  $\lambda$  value is difficult because both the trade-off property and the budget criterion should be considered simultaneously. As soon as the  $\lambda$  value is decided, the best operating point choice is easily made by minimization of the Lagrangian cost function.

### 3.3. Analysis of Source Coding Information Theory

Some issues when applying information theory to source coding are first discussed in this section. The application of the rate distortion function of wavelet coder with different content and coding parameters is then presented. Finally, the operations and insights of applying Lagrange Multiplier optimization techniques to source coding are described.

#### 3.3.1. Rate Distortion Function

In the previous section, the general form of rate distortion function is formulated

as in Eq. (10).

$$R_L(D) = \frac{1}{2} \log_2 \frac{\omega}{D}, \quad \omega = \frac{2^{2E(b)}}{2\pi e}. \quad (10)$$

Some literatures apply the function to embedded wavelet coder [5] [6], and make a little empirical adjustment on the parameters. The revised relationship with an additional parameter,  $\chi$ , is shown in Eq (11).

$$R(D) = \frac{1}{2} \chi \log_2 \frac{\omega}{D}, \quad \omega = \frac{2^{2E(b)}}{2\pi e}. \quad (11)$$

The parameter,  $\chi$ , characterizes the exponentially decaying rate. Base on the analysis of the experimental result in the literatures [5] [6], the parameter is proved to be related to the distribution of the source. As a result, the general rate distortion function which is suitable for embedded wavelet coder with square-error distortion measure is shown below.

$$\begin{aligned} R(D) &= \frac{1}{2 / \log_2 e} \chi \ln \frac{\omega}{D}, \quad \omega = \frac{2^{2E(b)}}{2\pi e}. \\ &= \gamma \ln \frac{\omega}{D} \end{aligned} \quad (12)$$

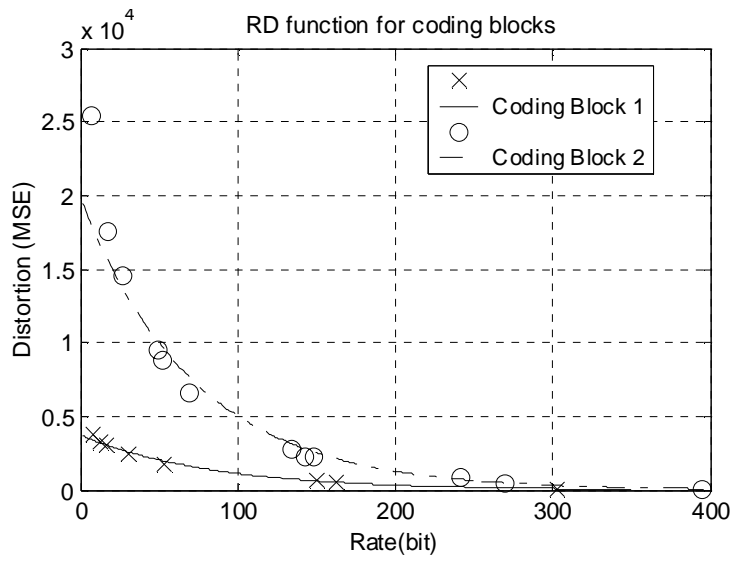
$$\text{or } D(R) = \omega e^{-\frac{R}{\gamma}}.$$

We conducted an experiment on the MSRA codec to examine the precision of the rate distortion relationship in Eq (12). The test sequence is Stefan in CIF resolution. The partial results for two coding blocks are shown in Fig 5. Each point in the figure represents an available truncated point in a coding block, and each curve represents the characteristic model for a coding block. The models are calculated by solving the parameter  $\gamma$ ,  $\omega$  in Eq (12) using least-squares-error curve fitting method. According to

the distribution of the content, these two coding blocks have different values of the parameters shown in Eq (13). The experiment shows the precision and the reliability of the rate distortion function when applying to coding blocks with different characteristics.

$$\text{Coding Block 1: } D_1(R_1) = 3739.1 e^{-0.012 R_1} . \tag{13}$$

$$\text{Coding Block 2: } D_2(R_2) = 19794 e^{-0.0137 R_2} .$$



**Fig 5. R-D Model for Coding Blocks in the MSRA Codec**

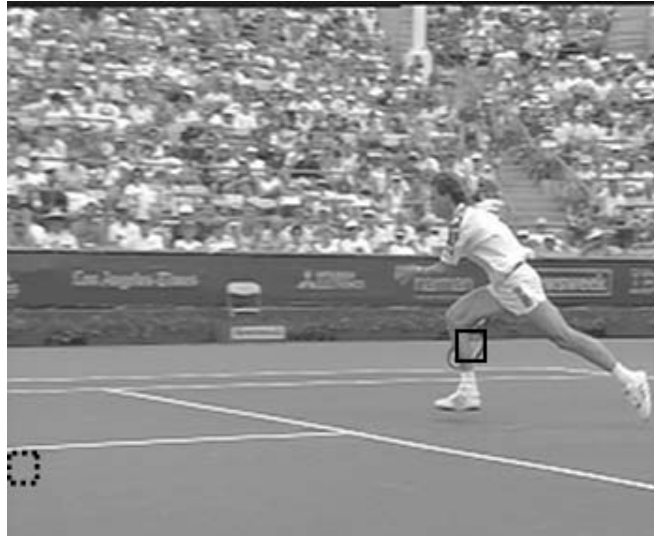
### 3.3.2. Lagrange Multiplier Optimization

The operations while applying the Lagrange Multiplier optimization technique in source coding are described as follows. For the constraint  $R \leq R_{target}$ , the “achievable” R-D point on the R-D curve with minimum distortion is the optimal solution. To solve the constrained optimization problem, it is easier to transform the problem into an unconstrained problem by adopting Lagrange optimization technique, as Eq (9). For a given Lagrange multiplier  $\lambda$ , a rate-distortion point with minimum  $J(R)$  in Eq (9) is the optimal solution.

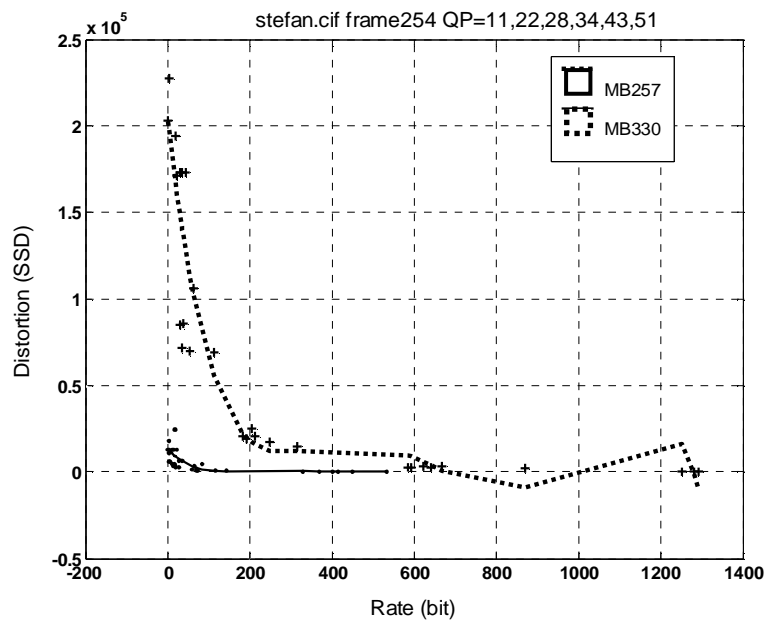
To demonstrates how R-D optimization works in practical applications, we conducted an

experiment using the JM reference software of the H.264/AVC codec. The R-D curves of possible coding modes corresponding to two macroblocks in Fig 6 are shown in Fig 7. Each curve in Fig 7 is generated with five different quantization parameters, and the coding modes including INTER-16x16, INTER-16x8, INTER-8x16, INTER-8x8, INTRA-16x16 and INTRA-4x4. Fig 8 and Fig 9 show magnified plots of the R-D curves corresponding to the curve of QP = 11 in Fig 7, and the slope of the tangent lines represent the value of the Lagrange multiplier. In addition, the arrows in the figures show the points the Lagrange multiplier line first hit, which are the optimized modes with the lowest value of the cost function.

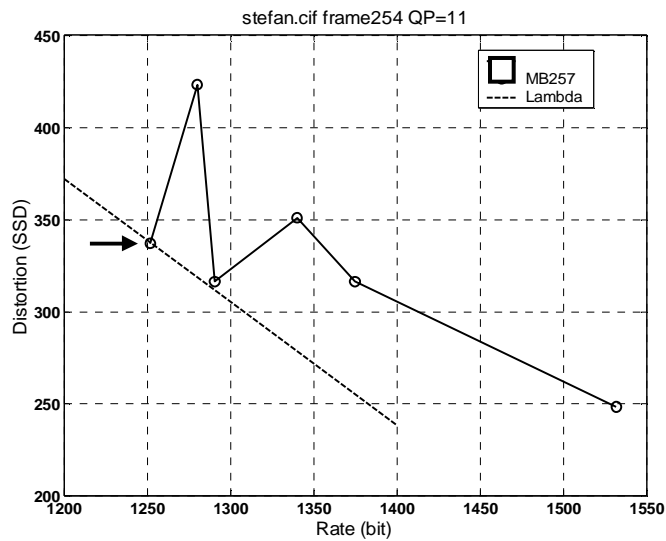
There are two insights to the R-D optimization scheme. First, for each coding unit, namely a macroblock in this example, R-D optimization mechanism makes coding decision according to the trade-off ratio, Lagrange multiplier. The rate distortion points which do not locate on the convex hull of all possible operating points will not be selected regardless of the Lagrange multiplier. Secondly, by applying the same Lagrange multiplier value to all macroblocks, the rate distortion optimization mechanism automatically distributes more bits to the macroblocks with more detail information because the rate distortion curves of these macroblocks are much steeper. In Fig 8 and Fig 9, the macroblock on Stefan's foot which comparably contains more complicated details allocates 1250 bits while the macroblock on the tennis court allocates only 400 bits.



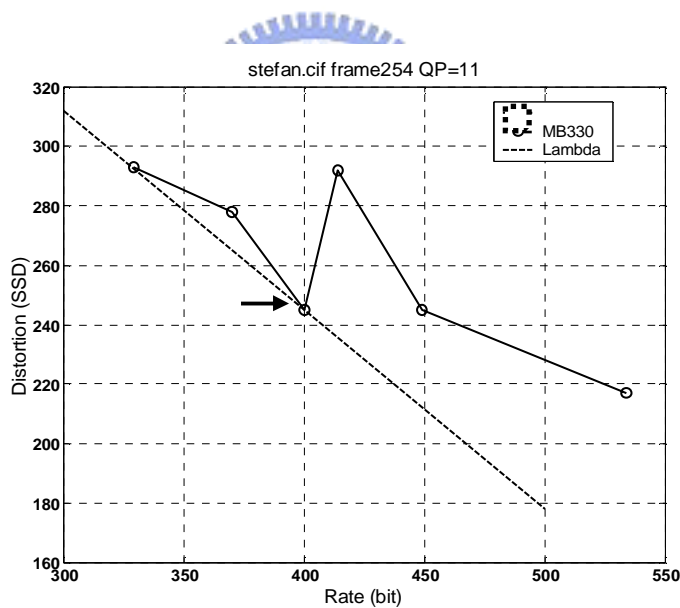
**Fig 6. Frame 254 in the video sequence Stefan**



**Fig 7. The RD curves corresponding to two MBs in Fig 6.**



**Fig 8. The enlarged RD curve around the optimal mode in Fig 7 for the MB on Stefan's foot.**



**Fig 9. The enlarged RD curve around the optimal mode in Fig 7 for the MB on the tennis court.**

### 3.4. Implementation Issues

So far, we have introduced the theory background of scalable rate control algorithms. However, there are still some gap between the theory and an actual



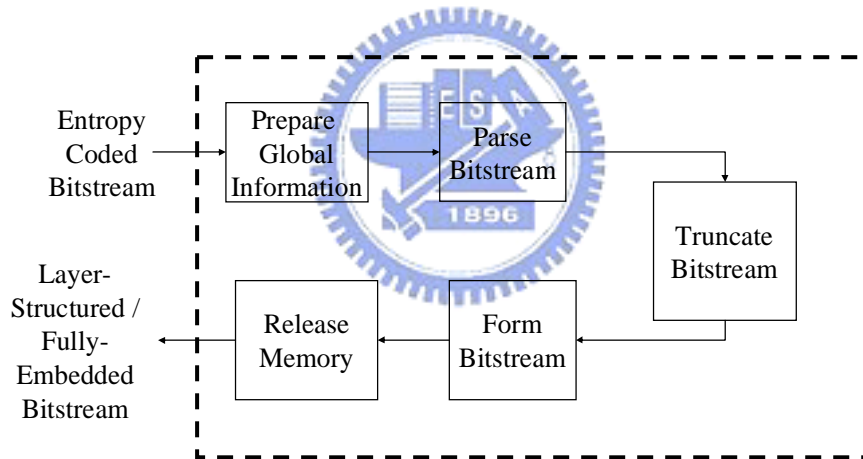
implementation. For example, the determination of the Lagrange multiplier value is difficult in practice, and the overall bit allocation procedure should be restructured in order to achieve computational efficiency. Solutions to these issues will be developed in the proposed scheme in the following chapters.



## 4. Proposed Rate Control Framework

In order to design a highly efficient rate control mechanism for resource constrained system, the general scheme for rate control extractor and the critical part in the current method are first clarified. Then the proposed algorithm including rate-lambda model and overall procedures are elaborated according to the theory basis introduced in the previous section. Finally, a particular application scenario, multiple-adaptation for media delivery, is introduced. The advantage of using the proposed algorithm in the multiple-adaptation scenario will be described.

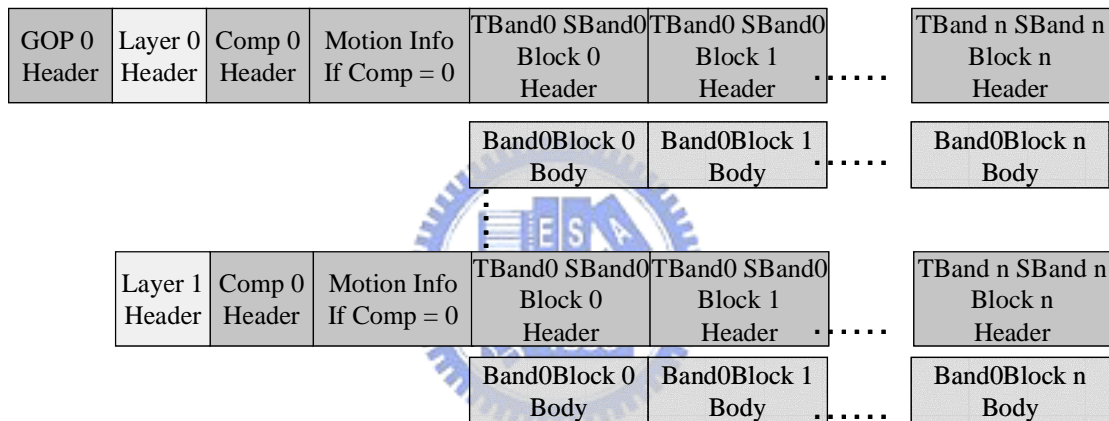
### 4.1. General Rate Control Extractor



**Fig 10. Flow Chart of MSRA Rate Control Extractor**

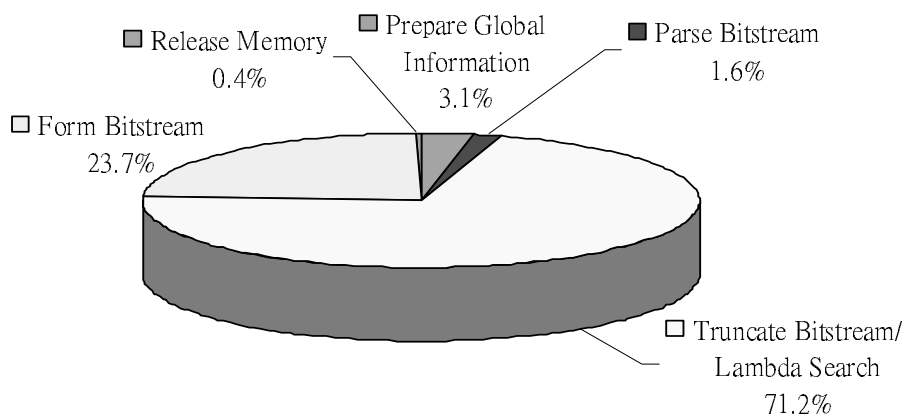
The general scheme for a rate control extractor is shown in the flow chart in Fig 10. The extractor may be executed on the server side right after the encoding procedure or be used as a stand-alone module on a transport gateway or even the client side. No matter which case the extractor belongs to, the general procedures begin with global information preparation. The global data includes the entropy-coded bitstream and the usage scenario criteria for the output bitstream. The usage scenario determines the video parameters such as the number of layers, the resolution,

the frame rate, and the bit rate for the target applications. A bitstream encoded using the MSRA codec is organized in the format shown in Fig 11. A bitstream parser extracts the information for the truncated candidates from the headers. After all the required data are collected, the bitstream truncation procedure begins without entropy decoding involved. The truncation module decides the truncation point in order to meet the resolution, frame rate, and bit rate criterions. The bitstream is then composed again with new header information and truncated body bits. The new bitstream should conform to the usage scenario and can be transmitted over the network to the target recipient.



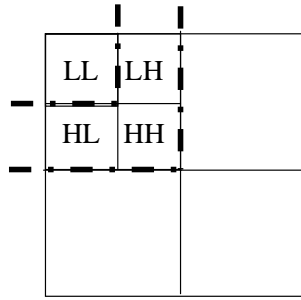
**Fig 11. Bitstream Format**

In order to reduce the computational cost of the rate control extractor, the profile of the computation time should be analyzed first. Fig 12 shows the percentage of computation time for each module in MSRA codec rate control extractor. The profiling is done by Microsoft Visual C++ development environment. The test sequence is Foreman in CIF resolution. The pie chart shows that the critical part for doing rate control is the truncation module. Therefore, reducing the complexity of bitstream truncation is crucial for a highly efficient rate control extractor.

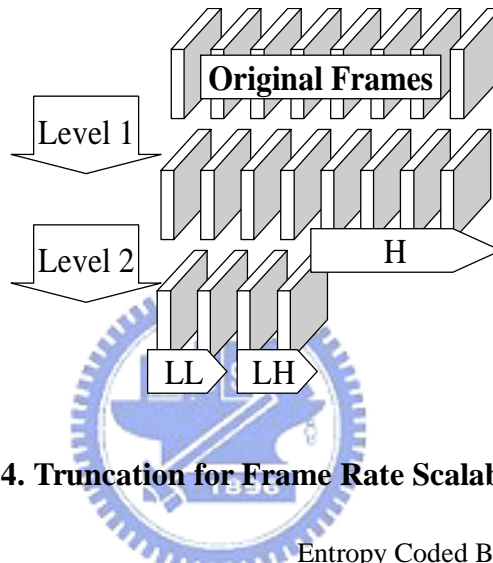


**Fig 12. Overall Computation Cost Analysis**

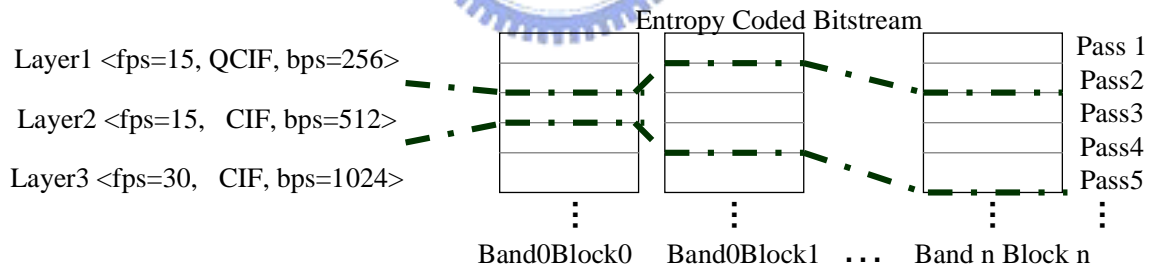
The truncation operation of an embedded scalable bitstream can be classified into three cases, including resolution change, frame rate change, and bitrate change. These cases are shown in Fig 13 to Fig 15. For resolution change, the truncation candidates are the boundaries of the spatial wavelet subband. The number of the candidates depends on the level of spatial wavelet decomposition. Similarly, temporal wavelet subbands contribute to the frame rate scalability. The low frequency band in the highest level has the highest priority. As the subband in the bitstream increases by one level, the frame rate is doubled. The bitrate scalability is the most complicated truncation procedure among the three cases. The candidate truncation points are the fractional pass in each bitplane for each coding block. The truncated bitplane passes for different coding blocks may be different according to the characteristics of the contents in the blocks. It is important to distribute the bits to visually more important subbands in order to compose a bitstream with best quality. The resulting bitstream is called a rate distortion optimized bitstream for a given target rate. Because the complexity of the bitrate scalability truncation procedures is very high, reducing its complexity is the main target of this thesis.



**Fig 13. Truncation for Resolution Scalability.**



**Fig 14. Truncation for Frame Rate Scalability.**

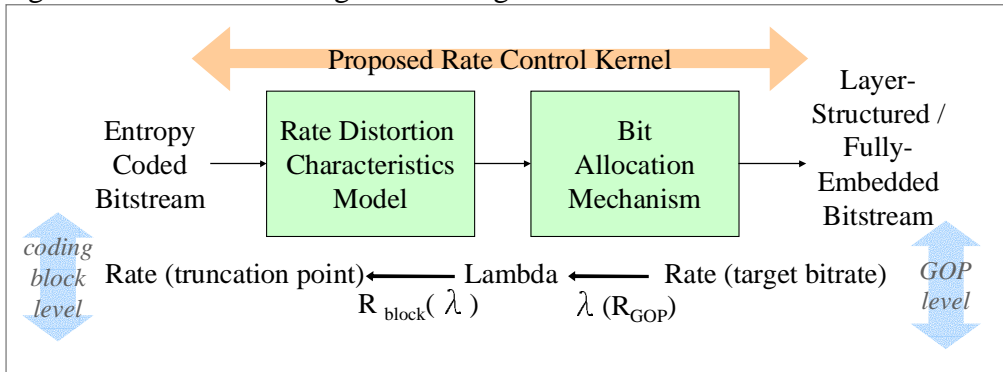


**Fig 15. Truncation for Bitrate (Quality) Scalability.**

## 4.2. Proposed Rate Control Extractor

According to the profiling results in the previous section, it is obvious that the truncation procedure for bitrate scalability is the most critical part in rate control scheme. Therefore, under the general rate control framework shown in Fig 16, the concept of the propose method tries to build a continuous rate lambda relationship

function for each coding block and each GOP in order to speed up the search for the optimal truncation points. The rate of each coding block corresponds to the truncation point, and the rate of each GOP corresponds to the target bit rate. These two values are related to each others by the lambda value. Therefore, the truncated point for each coding block can be selected given the target bit rate.



**Fig 16. The Concept of the Proposed Rate Control Extractor**

In the following subsections, the key technique of the thesis, the R-Lambda models for coding block level and GOP level, will first be introduced. Then the overall algorithm adopting the R-Lambda model will be elaborated. Finally, the comparison between the proposed method and the current approach will be presented.

#### 4.2.1. R-Lambda Model Analysis

The proposed R-Lambda model for each coding block is established by combining rate distortion function and Lagrange multiplier optimization technique. The rate distortion theory has been introduced in 3.2.1 and a practical example for wavelet based scalable video coding was illustrated in 3.3.1. The rate distortion function is repeated again in Eq (14) and the Lagrange function introduced in 3.2.2 is shown in Eq (15).

$$R(D) = \gamma \ln \frac{\omega}{D}. \quad (14)$$

$$J(R) = D + \lambda R. \quad (15)$$

In Eq (14), the parameter  $\gamma$  depends on the distribution of the source, and the parameter  $\omega$  is related to signal variance. For a given value  $\lambda$ , the minimization of  $J(R)$  in Eq (15) can be obtained when the first derivative  $dJ(R)/dR = 0$ , that is,

$$\frac{dJ(R)}{dR} = \frac{dD(R)}{dR} + \lambda = 0, \text{ and} \quad (16)$$

$$\lambda = -\frac{dD(R)}{dR}. \quad (17)$$

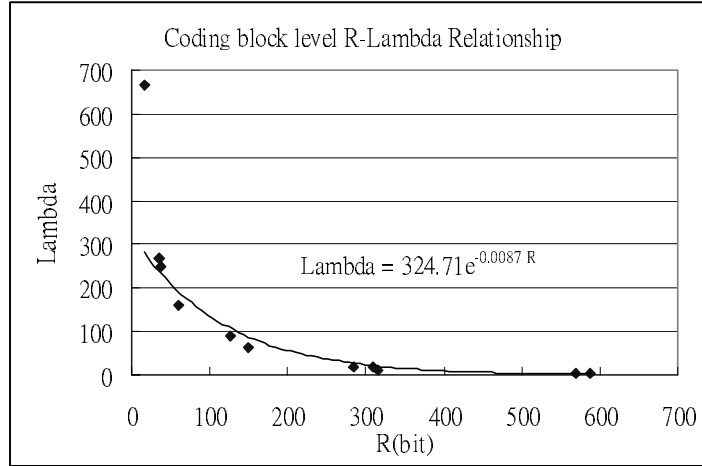
Solve Eq (14) for  $D$  and substitute it into Eq (17), the relationship between the Lagrange multiplier and the bitrate can be derived as follows,

$$\begin{aligned} \lambda &= -\frac{dD(R)}{dR} = -\frac{d(R^{-1}(D))}{dR} \\ &= -\frac{d\left(\alpha e^{-R/\gamma}\right)}{dR} = \left(\frac{1}{\gamma}\right)\omega e^{-R/\gamma} \end{aligned} \quad (18)$$

In summary, the R-Lambda model in coding block level can be written as in Eq (19) where the parameters  $\alpha$  and  $\beta$  are source dependent:

$$\lambda = \alpha e^{\beta R}. \quad (19)$$

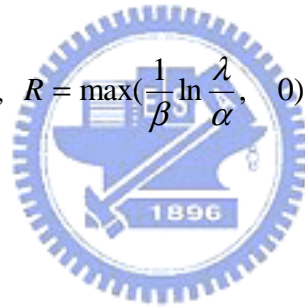
In order to prove the accuracy of the model, an experiment is conducted on MSRA codec. Fig 17 illustrates the practical situation for the test sequence, FOOTBALL. Each point in the graph represents a truncation point. By curve fitting, the RD information for each truncated point can be represented as the equation with only two parameters.



**Fig 17. Coding Block level R-Lambda Relationship Examination**

The GOP level R-lambda model can be extended from the model in coding block level as in Eq (20) and Eq. (21):

$$\lambda = \alpha e^{\beta R}, \quad R = \max\left(\frac{1}{\beta} \ln \frac{\lambda}{\alpha}, 0\right), \quad (20)$$



For  $\alpha > 0, \beta < 0$

$$\begin{aligned} R_{GOP} &= \sum_i R_{block\ i} = \sum_i \max\left(\frac{1}{\beta_i} \ln \frac{\lambda}{\alpha_i}, 0\right) \\ &= \sum_j \frac{1}{\beta_j} \ln \frac{\lambda}{\alpha_j} \quad \text{which } \{j \in S \mid \alpha_j > \lambda \text{ in } S\} \\ &= \sum_j \frac{1}{\beta_j} (\ln \lambda - \ln \alpha_j) \\ &= \left(\sum_j \frac{1}{\beta_j}\right) \ln \lambda - \left(\sum_j \frac{1}{\beta_j} \ln \alpha_j\right) \end{aligned} \quad (21)$$

It is straightforward that the rate in a GOP is the sum of the rate in a group of coding blocks (Eq (21)), and the size of the group is related to the lambda value. and therefore the two summation terms P and Q as defined in Eq.(22):



$$p_{GOP} = \sum_j \frac{1}{\beta_j} \quad \text{and} \quad q_{GOP} = \sum_j \frac{1}{\beta_j} \ln \alpha_j, \quad (22)$$

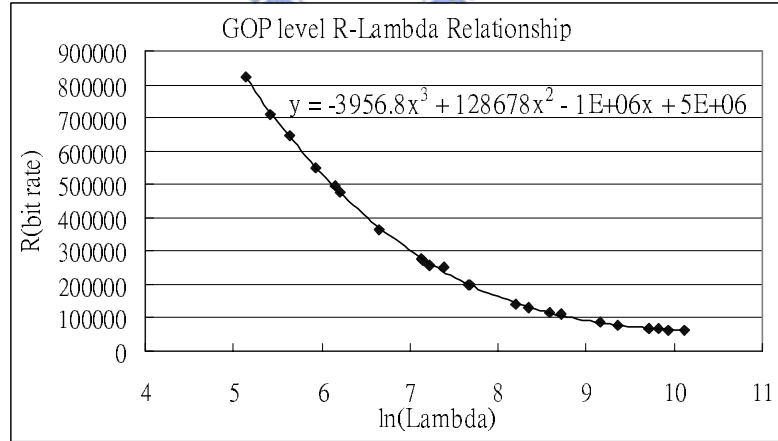
In order to keep the model simple, we assume that these two summations can be modeled by polynomials (Eq (23)) as follows:

$$P_{GOP} = a_1(\ln(\lambda))^{n-1} + a_2(\ln(\lambda))^{n-2} + \dots + a_n \quad \text{and} \quad (23)$$

$$q_{GOP} = b_1(\ln(\lambda))^{n-1} + b_2(\ln(\lambda))^{n-2} + \dots + b_n.$$

Finally, the relationship of the GOP level R-lambda model is established (Eq. (24)).

$$R_{GOP} = p_{GOP} \ln \lambda - q_{GOP} = \gamma_1(\ln \lambda)^n + \gamma_2(\ln \lambda)^{n-1} + \dots + \gamma_{n+1} \quad (24)$$



**Fig 18. GOP level R-Lambda Relationship Examination**

The graph in Fig 18 illustrates the accuracy of the proposed R-lambda model in the GOP level. The order of the function is determined by the experience of the experimental results. Statistically, a cubic function can be used to fit the data points

well for a wide range of rate.

#### 4.2.2. Overall Framework

The proposed algorithm uses the R-lambda model for two purposes. One is to search for the optimal  $\lambda$  for a quality layer in the GOP level, and the other is to describe the R-D characteristics of a single block in the coding block level.

The coding block level model Eq (19) is used as an adaptive model since the source dependent parameters  $\alpha$  and  $\beta$  are estimated causally based on the input data. Given  $n$  pairs of numerical data  $(\lambda_i, R_i)$ ,  $i = 0, \dots, n - 1$ , the parameter  $\alpha$  and  $\beta$  can be calculated as follows. First, Eq (19) can be rewritten as  $\ln\lambda = \ln\alpha + \beta \cdot R$ . Therefore, for  $n > 2$  we have an over-determined system of equations,

$$\begin{pmatrix} \ln \lambda_0 \\ \ln \lambda_1 \\ \vdots \\ \ln \lambda_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & R_0 \\ 1 & R_1 \\ \vdots & \vdots \\ 1 & R_{n-1} \end{pmatrix} \begin{pmatrix} \ln \alpha \\ \beta \end{pmatrix}. \quad (25)$$

The system can be solved by the least-squares estimation. Once the parameters  $\alpha$  and  $\beta$  are determined, the relationship between the Lagrange multiplier and rate is directly established. In a similar manner, the GOP level R-lambda model (Eq (24)) is adaptively built by the least-squares curve fitting method. For certain GOP, assume that

$$Y = \begin{pmatrix} R_{GOP1} \\ R_{GOP2} \\ \vdots \end{pmatrix}, \quad A = \begin{pmatrix} (\ln \lambda_1)^n & (\ln \lambda_1)^{n-1} \cdots & 1 \\ (\ln \lambda_2)^n & (\ln \lambda_2)^{n-1} \cdots & 1 \\ \vdots & \vdots & \vdots \end{pmatrix}, \quad \text{and} \quad X = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{n+1} \end{pmatrix}, \quad (26)$$

the parameters  $\gamma_1, \gamma_2, \dots, \gamma_n$  are solved by the matrix operations:

$$Y = AX, \text{ and}$$

(27)

$$X = (A^T A)^{-1} A^T Y .$$

As the whole GOP level R-lambda model is established, the lambda value is solved by the following algebraic functions.

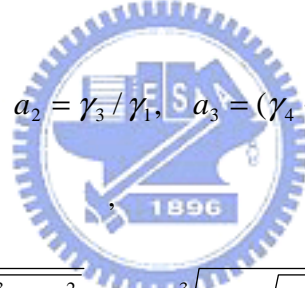
If n=2,

$$\ln \lambda = \frac{-\gamma_2 - \sqrt{\gamma_2^2 - 4\gamma_1(\gamma_3 - R_{target})}}{2\gamma_1} .$$

(28)

If n=3,

$$\text{let } a_1 = \gamma_2 / \gamma_1, \quad a_2 = \gamma_3 / \gamma_1, \quad a_3 = (\gamma_4 - R_{target}) / \gamma_1,$$

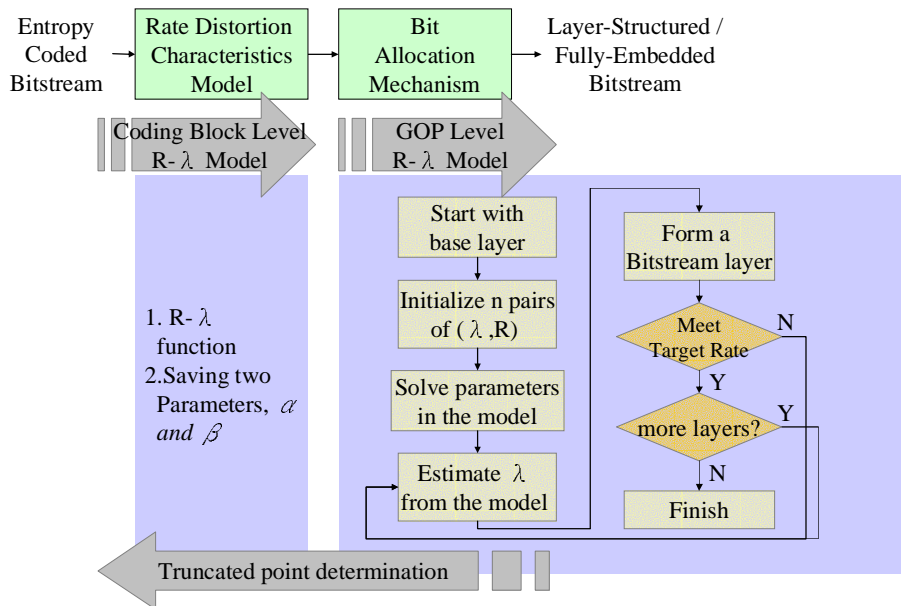


$$S = \sqrt[3]{N + \sqrt{M^3 + N^2}}, \quad T = \sqrt[3]{N - \sqrt{M^3 + N^2}}$$

(29)

$$\Rightarrow \ln \lambda = \begin{cases} S + T - \frac{1}{3}a_1 \\ -\frac{1}{2}(S + T) - \frac{1}{3}a_1 + \frac{1}{2}i\sqrt{3}(S - T) \\ -\frac{1}{2}(S + T) - \frac{1}{3}a_1 - \frac{1}{2}i\sqrt{3}(S - T) \end{cases}$$

The overall proposed algorithm which adopts the R-lambda model is described as follows (Fig 19). In the bit allocation mechanism, the R-lambda model is used to search the lambda value in the GOP level, and in the rate distortion optimization procedure, the R-lambda function is used to represent the rate distortion properties in the coding block level.



**Fig 19. Overall Framework of the Proposed Rate Control Mechanism**

1、 Search for the optimal Lagrange multiplier : The GOP level R-lambda model is adopted in this step to simulate the behavior of data in a GOP. The aim is to speed up optimal Lagrange multiplier search by better understanding of the R-lambda relationship rather than blind iterative search with bisection method. Detail flow chart of the algorithm is described step by step as follows:

- a) Find the first n pairs of  $(\lambda, R)$  in the base layer, and n is typically 4 for the cubic model in GOP level.
- b) Solve for the parameter  $(\gamma_1, \gamma_2, \dots, \gamma_3)$  using Eq (27).
- c) Given target bitrate, estimate  $\lambda$  using Eq (29).
- d) Use the estimated  $\lambda$  to form the bitstream layer and obtain another  $(\lambda, R)$  data point.
- e) Add the new  $(\lambda, R)$  pair to the data set.
- f) Iteratively doing b)-e) until the R value is close enough to the target bitrate within a tolerable error range TR.
- h) Repeat the procedure for the enhancement layers.

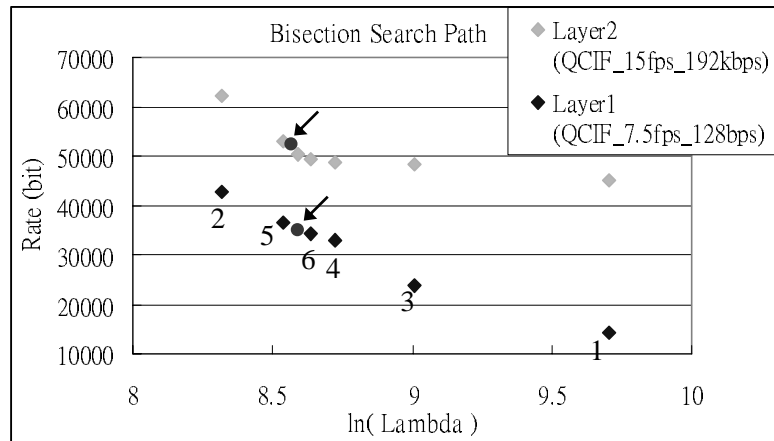
In the proposed algorithm, no additional memory storage is needed. Furthermore, the smaller the  $n$  is, the lower the computational overhead will be. The experimental results in next section show that even for a small  $n$  value, the accuracy of the algorithm is still good.

2、 Represent RD property of a coding block: In procedure d), a bitstream layer is formed given a Lagrange multiplier value. The truncation point of each coding block is determined at the fractional bitplane pass with the nearest Lagrange multiplier value. To achieve the typical coding block level rate allocation, the Lagrange multiplier value of each fractional bitplane pass in all coding blocks should be stored during tier 1 of entropy coding. In order to reduce the memory usage of the information and distribute the rate among all coding blocks based on information theory, the coding block level R-lambda model is applied to describe the property of each coding block. Therefore, only the parameters  $\alpha$  and  $\beta$  should be stored for a single coding block, and the coding block level rate allocation can be easily done by adopting the inverse R-lambda model with a given Lagrange multiplier. In the proposed method, the truncation point would be the fractional bitplane pass with the nearest rate.

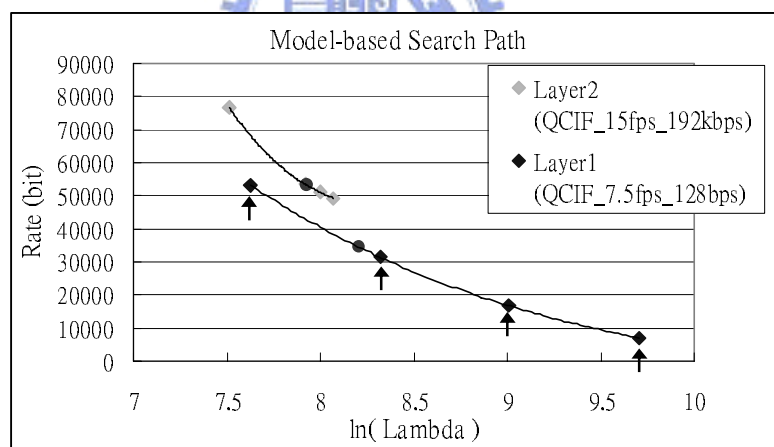
### **4.2.3. Lambda Search Procedure**

A real example of the comparison of different lambda search procedures in the GOP level are shown in the graphs Fig 20, Fig 21. The experiment is conducted using the MSRA codec, the test sequence is FOOTBALL in CIF resolution with frame rate 30 frames per second. The truncation criterions are QCIF resolution, 7.5 frames per second, bitrate 128 bps for layer 1 and QCIF resolution, 15 frames per second, bitrate 192 bps for layer 2. The original codec uses the conventional bisection method. The search approach starts from the initial maximum and minimum lambda estimation. After the bitstream formation, the exact bitrate value will be obtained. Then the next

estimation of the lambda value is at half of the range between maximum and minimum values. By half-eliminating the search range at each iterative operation, the search results converge and the lambda point which meets the target bitrate is obtained at the end. The number in Fig 20 shows the path of each searching step.



**Fig 20. Search Path for Bisection Method**

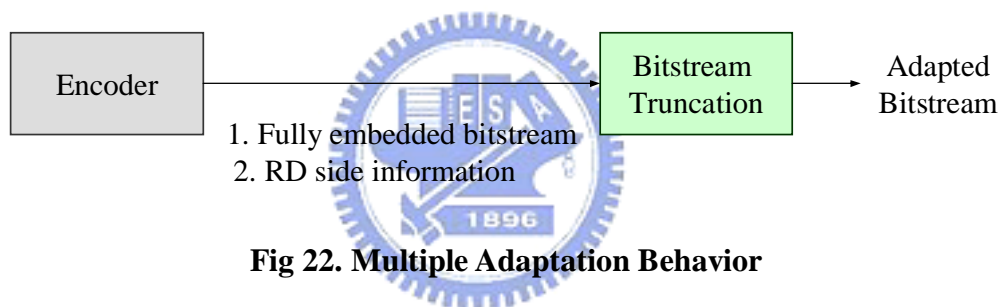


**Fig 21. Search Path for the Proposed Method**

For using the proposed algorithm, the lambda value is estimated in a different way. Because the model is implemented with a cubic function, the procedures should start from four initial guess at least which the arrows point to in Fig 21. Then the model is form using curve fitting method in the figure. Usually, the lambda estimation using the proposed model can meet the target in two steps.

The basic difference between these two methods is in the estimation of the rate distortion behavior. The previous method uses blind bisection search to decide the bitstream formulation method only by using the real data points. The rate distortion behavior can be obtained when the test points achieve certain quantity. On the other hand, the proposed approach tries to model the form of the rate distortion behavior at the beginning by using the  $R-\lambda$  model, and fine tune the parameters of the model according to the real data points. Therefore, the model represents the characteristic of the content well, the accuracy of the predicted  $\lambda$  value is higher, and the converging speed is faster.

### 4.3. Proposed Multiple Adaptation Scheme



**Fig 22. Multiple Adaptation Behavior**

In a multiple adaptation system, the sending device should provide both a fully embedded bitstream and the RD side information to the receiving device in order for it to do the second truncation. The bitstream can be adapted and transmitted several times to different devices with different channel bandwidth and capabilities. The bitstream truncation can be executed several times without complete bitstream decoding involved. A multiple adaptation scheme is useful only when the system overhead is not too high. That is, the RD side information is small and the complexity of the truncation module is low.

The side information in the MSRA codec is the discrete rate-lambda pairs for all coding passes. On the contrary, the proposed method translates the discrete R-D values into a close form R-lambda function with only two parameters. As a result, the

size of the side information is reduced. Based on the computation cost analysis from the previous section, the number of iterations of the search procedures also reduces. More experimental results will be shown in chapter 5 to support the advantages of the proposed model.





## 5. Experimental Results

In this section, some experiments on the proposed algorithm are conducted using the MSRA scalable video codec, with the MPEG test sequences, STEFAN, FOREMAN, MOBILE and FOOTBALL in CIF resolution. The coding parameters used in the experiments are as follows. The GOP size is 64 frames, and the frame rate is 30 fps. The parameter  $n$  in the GOP level model is set to 3, and the bitrate error threshold  $T_R$  is set to 3% of the target bitrate.

Even though the experiments are conducted on a wavelet-based scalable video codec, similar results should apply to wavelet-based still image codec due to the similarity of these techniques.

### 5.1. Computational Cost Reduction in Rate Control Extractor

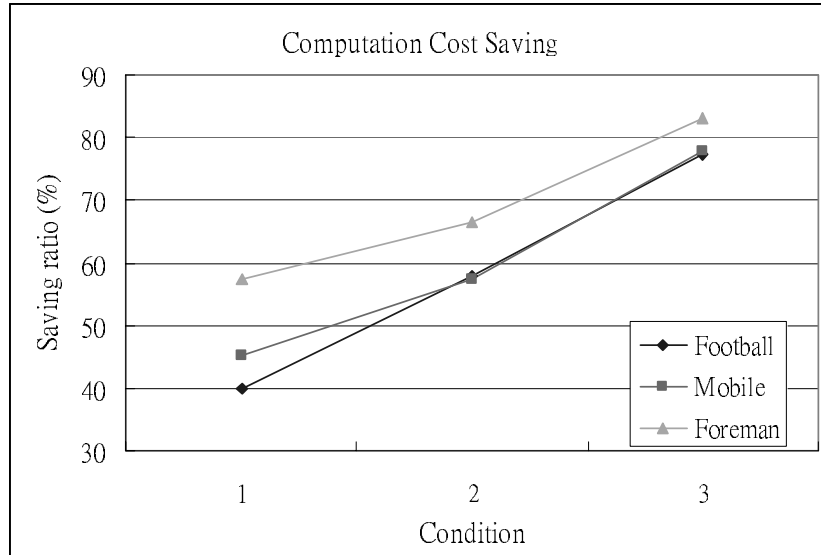
The number of iterations required before the solution converges for the proposed method and the commonly used bisection search is shown in TABLE I. The average computation cost saving is about 47% when the resolution and frame rate setting for each layer are all different.

<i>Sequence</i>	<i>MSRA Bisection</i>	<i>R- <math>\lambda</math> Model</i>	<i>Saving Ratio</i>
<b>Mobile</b>	9.67	5.30	45.17 %
<b>Foreman</b>	10,68	4.55	57.41 %
<b>Football</b>	7.84	4.70	40.05 %
<b>Average</b>	<b>9.40</b>	<b>4.85</b>	<b>47.54 %</b>

**Table 1. Number of Iterations Comparison for Lambda Search**

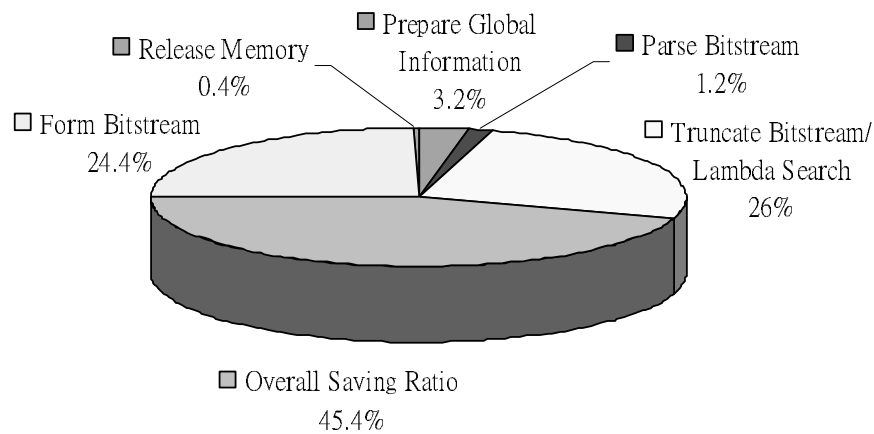
When the number of layers for the each resolution and frame rate setting

increases, the search procedure can converge even faster by taking advantage of the R-lambda model from the previous layer. According to our experiments, the saving ratio is about 60% when the layer number is 5, and up to 80% when the layer number is 12 (Fig 23).



**Fig 23. Saving Ratio of the Iteration Times**

Proposed Model-based Rate Control Extractor

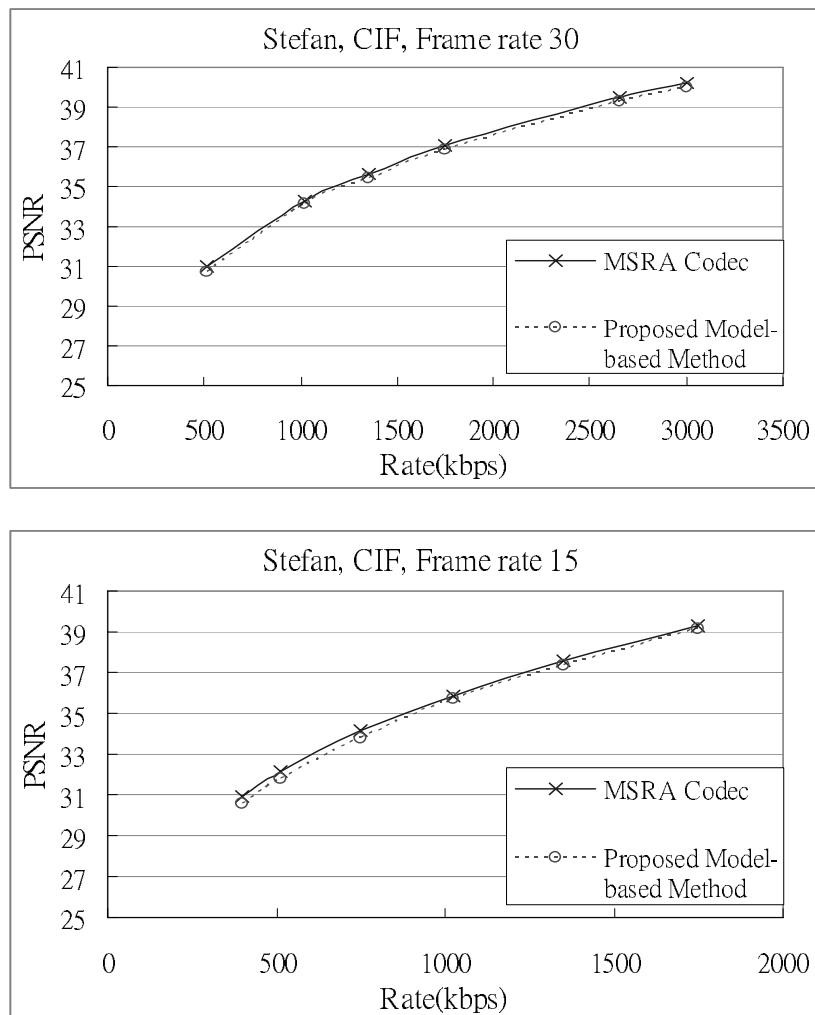


**Fig 24. Overall Computational Cost Saving for Rate Control Extractor**

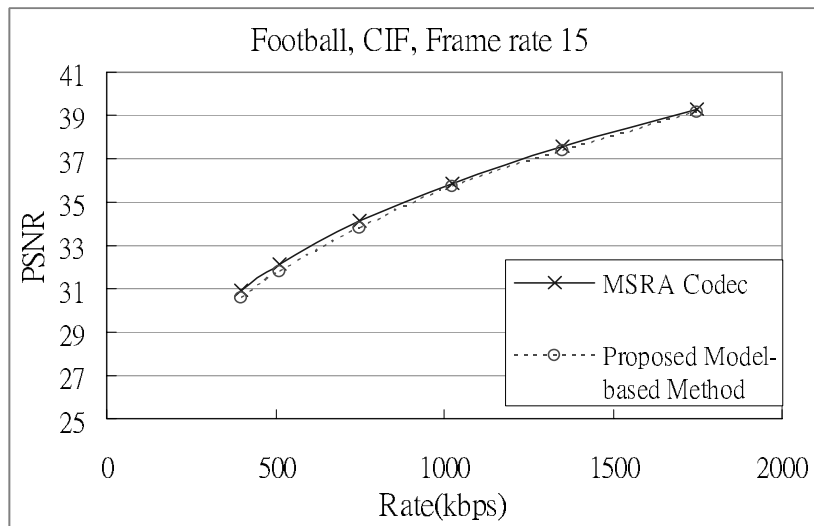
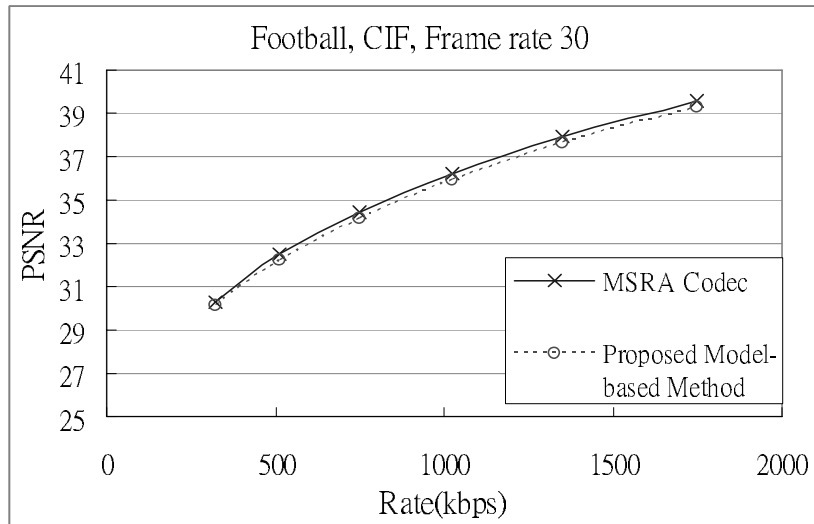
Besides the lambda search procedures, the overall computational cost saving is shown in Fig 24. The test sequence is FOREMAN in CIF resolution and the rate control extractor truncates the bitstream from 768 bps to 256bps. The CPU usage

percentage of the bitstream truncation module is reduced from 71% to 26%. Namely, the overall saving ratio is about 45%.

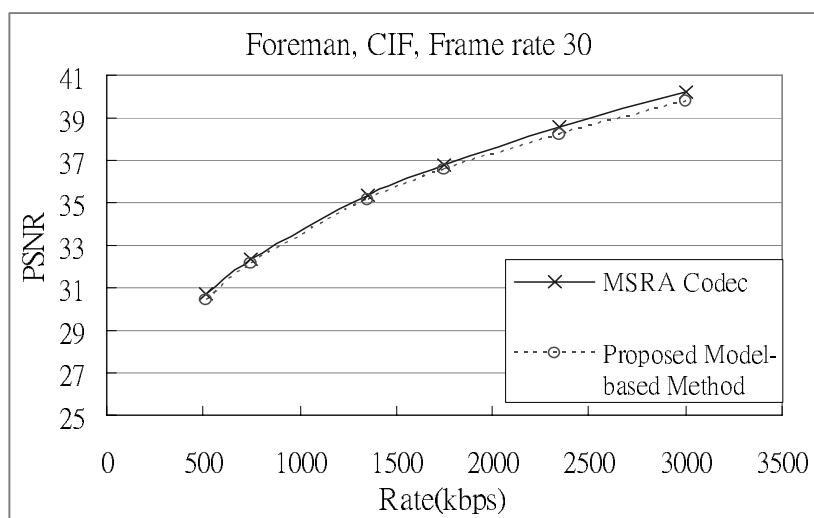
Since the proposed mechanism allocates rate for each coding block in a model based method, the rate distribution in a GOP is rearranged. The coding efficiency graph is shown in Fig 25, Fig 26 and Fig 27. The test sequences are STEFAN, MOBILE, and FOREMAN in CIF resolution and are truncated at frame rate 30 and 15. The figures show that the proposed rate control mechanism achieves similar PSNR performance with MSRA codec at any range of the rate. The average PSNR degradation is only 0.25dB.

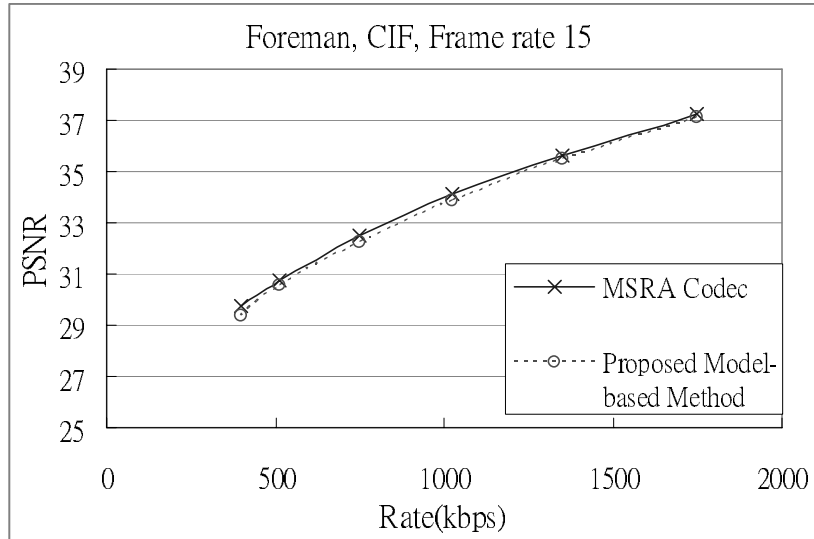


**Fig 25. PSNR Performance Comparison of Stefan**



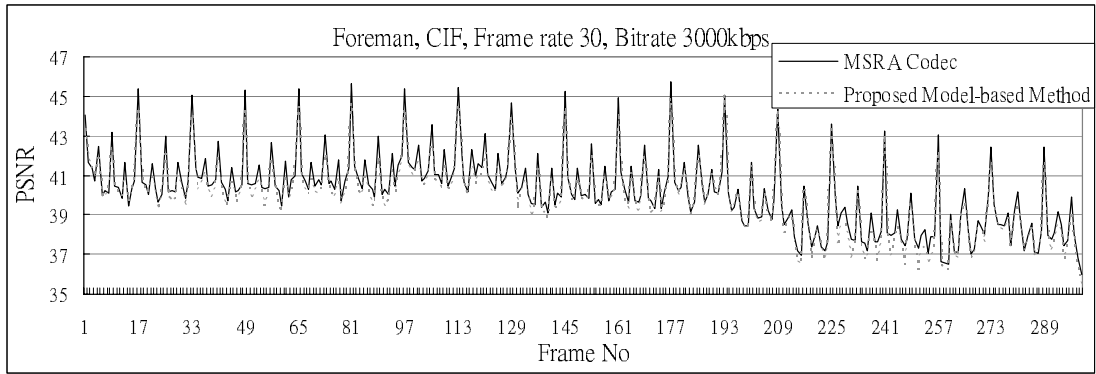
**Fig 26. PSNR Performance Comparison of Football**



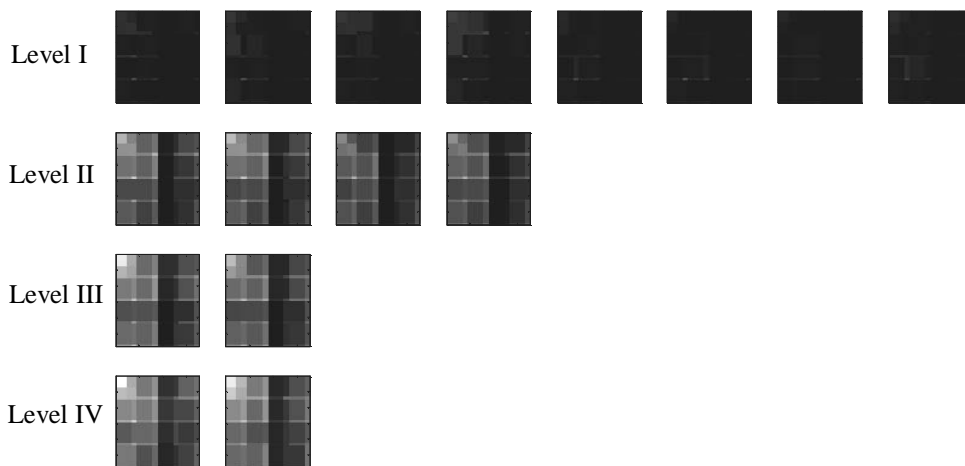


**Fig 27. PSNR Performance Comparison of Foreman**

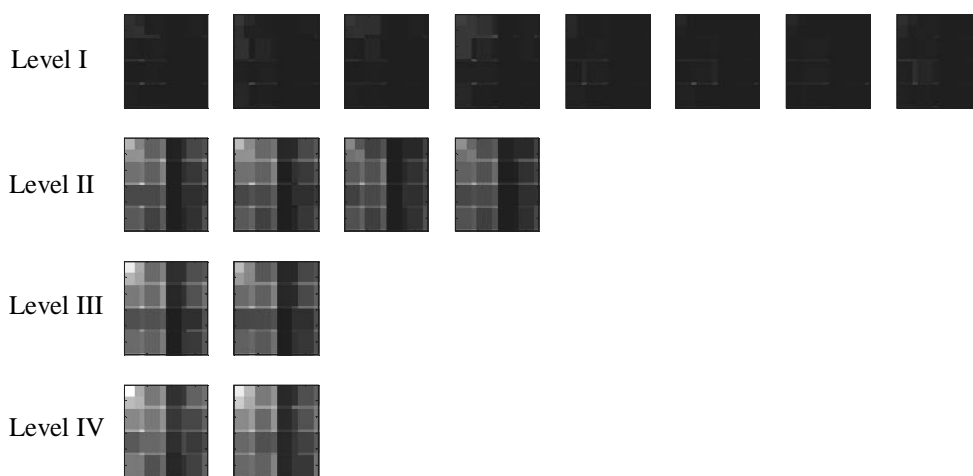
The reason for the PSNR degradation of the proposed approach is explained as follows. Fig 28 shows the PSNR performance comparison of all frames in the FOREMAN sequence. The frame level PSNR value of the proposed method meets the value of original method well with a little degradation in each frame. Taking the second GOP for example, the results of bit allocation is show in Fig 29 - Fig 31. The frames in the first row represent the highest frequency in temporal direction, and the top-left blocks in each frame represent the lowest frequency in spatial direction. The comparison between Fig 29 and Fig 30 is shown in Fig 31. The blocks with white color get more bits by using the proposed approach, and the blocks with black color obtain fewer bits. The figure shows that the proposed method allocates fewer bits to the top-left blocks in the low frequency frames which contain the most important information. As a result, the PSNR performance in the proposed method suffers a little degradation.



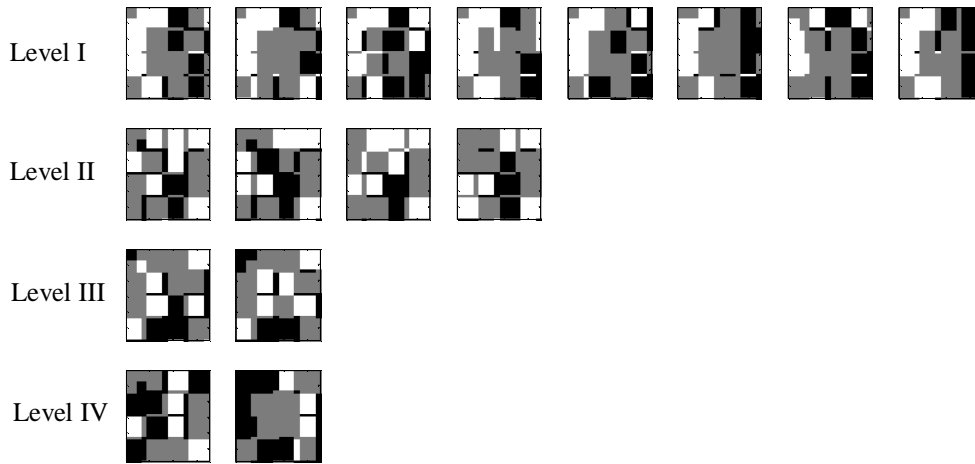
**Fig 28. Frame Level PSNR Performance Comparison of Foreman**



**Fig 29. Bit Allocation of the Second GOP in Foreman with MSRA Codec**



**Fig 30. Bit Allocation of the Second GOP in Foreman with the Proposed Method**



**Fig 31. Bit Allocation Comparison of the Second GOP in Foreman**

## 5.2. Side Information Saving for Multiple Adaptation Scheme

The experimental result in the Table 2 shows the saving ratio in different resolution and frame rate for different sequences in a multiple-adaptation scenario. The average saving ratio is about 50%, and the side information percentage in the bitstream is reduced from 3.5% to 1.7%.

<i>Sequence Name</i>	<i>Resolution / frame rate / bit rate (kbps)</i>	<i>Side information bits (% in bitstream)</i>		<i>Saving Ratio</i>
		<i>MSRA</i>	<i>Proposed Method</i>	
<b>Mobile</b>	CIF / 30 / 768	266,864 (3.39%)	128,475 (1.63%)	51.86%
	CIF / 30 / 512	165,691 (3.16%)	78,030 (1.49%)	52.91%
	CIF / 30 / 384	112,793 (2.87%)	52,924 (1.35%)	53.08%
	CIF / 15 / 256	73,981 (2.82%)	32,718 (1.26%)	55.78%
<b>Foreman</b>	CIF / 30 / 768	332,417 (4.23%)	166,249 (2.12%)	49.99%
	CIF / 30 / 512	234,343 (4.47 %)	107,864 (2.06%)	53.97%
	CIF / 30 / 256	109,067 (4.17%)	49,737 (1.90%)	54.40%
	CIF / 15 / 192	86,234 (4.40 %)	41,488 (2.12%)	51.89%
<b>Football</b>	CIF / 30 / 1380	347,746 (3.06%)	186,822 (1.64%)	46.28%
	CIF / 30 / 1024	282,381 (3.11%)	149,396 (1.64%)	47.09%
	CIF / 30 / 768	213,047 (3.13%)	110,072 (1.61%)	48.33%
	CIF / 15 / 512	141,205 (3.11%)	76,603 (1.69%)	45.75%
<b>Average</b>		3.5%	1.7%	50.94%

**Table 2. Side Information Saving Ratio**

## 6. Conclusion and Future Work

In this thesis, a novel adaptive model-based rate allocation mechanism for embedded wavelet image/video coding is proposed. By using the R-lambda model, a low complexity search procedures in the rate control mechanism is proposed. The proposed approach has many advantages over the existing ones. Detail comparisons of different rate control algorithms are listed in Table 3.

In the rate distortion characteristics module, the proposed method translates the behavior description from discrete data into a simple model. The coding block level  $\lambda_{cb}(\text{rate})$  model is prepared for GOP level bit allocation. The lambda value search procedure is achieved by using a adaptive  $\text{rate}_{\text{GOP}}(\lambda)$  model constructed at runtime instead of blind bisection search. And the distributed bit for each coding block is decided by the inverse  $\lambda_{cb}(\text{rate})$  model. The truncation pass is then determined by finding the pass with the nearest rate value. All these procedures aims at designing a low computational cost rate control extractor.

In addition, the R-D side information for multiple adaptation applications is reduced, which provides another advantage for the proposed method. The side information includes only two parameters in  $\lambda_{cb}(\text{rate})$  model for each coding block instead of all discrete rate-lambda pairs for all coding pass. Therefore, the size of the side information is reduced.



Rate Control Extractor Module		Original MSRA codec	Proposed Model-based approach	Advantage of the Proposed method
RD Characteristics / Behavior Description (for each coding block)		Discrete rate-lambda pairs for each pass	Lambda(rate <sub>cb</sub> ) model build	Speed up search procedure with low computational cost
Bit Allocation (for each GOP)	Lambda Search / Decision	Bisection Search	Adaptively build Rate <sub>GOP</sub> (lambda) model and estimate lambda value by the model	
	Truncation Pass Decision	Search coding pass with the nearest lambda value	<ol style="list-style-type: none"> <li>1. Obtain distributed R by inverse Lambda(rate<sub>cb</sub>) model in coding block level</li> <li>2. find the pass with the nearest rate value</li> </ol>	
Side Information		Discrete rate-lambda pairs	Two parameters in Lambda(rate <sub>cb</sub> ) model	Low side information requirement

**Table 3. Algorithm Comparison**

There are some differences in the design of the rate control mechanisms for different codecs. The comparisons are shown in Table 4. In the case of H.264, the output format is not an embedded bitstream. The rate control mechanism is involved in the encoding loop. Namely, the post-encoding rate control mechanism is not suitable for this type of codec. And fractional bit plane coding which can be easily used for rate distortion information extraction is not adopted. Therefore, the rate distortion behavior analysis can not be performed at runtime unless the encoding loop is executed several times with different quantization values. The approach increases the complexity of the system. A better way to design a rate control mechanism for a non-embedded bitstream is by buffer fullness analysis or content complexity analysis. These approaches can be applied efficiently in the encoding loop and the approximate rate distortion behavior can be obtained.

<b>Comparison</b>	H.264	Wavelet-based Scalable Codec of MSRA	JSVM (H.264 Scalable Extension of HHI)
Embedded bitstream	X	O	O
Fractional bit plane coding	X	O	O (Progressive refinement)
Rate Distortion Behavior Analysis	1. Rate distortion behavior is not easily built by runtime data. 2. Several encoding loops is needed by adopting different QP.	Rate distortion behavior can be built by using fraction bit plane information	Rate distortion behavior can be built by using progressive refinement information
Rate Control Module	Buffer fullness analysis and content complexity analysis	The proposed adaptive model can be used by collecting the rd behavior	The proposed adaptive model can be used by collecting the rd behavior

**Table 4. Rate Control Mechanism Comparison**

For embedded bitstream compressed by some scalable codecs, such as the wavelet-based codec by MSRA or JSVM by HHI, the architecture and bitstream format is suitable for post-encoding rate control mechanism. The rate distortion behavior can be built by using fractional bit plane information, which is called progressive refinement in JSVM. Here, the bit plane coding is divided into three scans in order to increase the rate distortion points on the convex hull and also increase the truncated points. By taking advantage of this feature, the overall rate distortion behavior is easily obtained and can be described by the proposed adaptive model. Hence, the rate control mechanism is more efficient than the methods which use trial and error approaches to achieve the target bitrate. For embedded coder with this feature, the rate distortion can be model by a function and the proposed model can be adopted as an accelerator.

There are still quite some improvements that can be made to the proposed

algorithm. For example, one can refine the R-D model and introduce visual quality measures into the rate allocation process. In addition, the motion bits should be taken into account and can be analyzed independently using a different model because the impacts of the motion information and the transform coefficients are different. The algorithm can be implemented on resource critical embedded system with further consideration of some platform issues. For example, for a hardware-based solution, the operations should be simplified further to follow the parallelism and data reused guidelines for hardware accelerator design. Further improvements can be expected with these efforts.



## 7. References

- [1] J. M. Shapiro, "An embedded hierarchical image coder using zerotrees of wavelet coefficient," IEEE Data Conference, Snowbird, UT 1993, pp. 214-223.
- [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 3D subband coding of video," IEEE Transaction on Image Processing, Vol. 3, pp. 572-588, Sep. 1994.
- [4] D. Taubman, "High performance scalable image compression with EBCOT," Image Processing, IEEE transactions on, vol. 9, pp. 1158-1170, July 2000.
- [5] Jin Li and C.-C. Jay Kuo, "Embedded wavelet packet image coder with fast rate-distortion optimized decomposition," Proc. SPIE: Visual Communications and Image Processing'97, Vol. 3024, pp. 1077-1088, 1997.
- [6] Po-Yuen Cheng, Jin Li, and C.-C. Jay Kuo, "Rate control for and embedded wavelet video coder," IEEE Transaction on Circuit and System Video Technology, vol. 7, NO. 4, Aug. 1997.
- [7] A. Aminlou and O. Fatemi, "Very Fast Bit Allocation Algorithm, Based on Simplified R-D Curve Modeling," Proceedings of 10th IEEE International Conferences on Electronics, Circuits, and Systems 2003, pp. 112-115, Dec. 2003.
- [8] J. Xu, Z. Xiong, S. Li and Y.Q. Zhang, "Three-Dimensional embedded subband coding with optimal truncation (3-D ESCOT)," Applied and Computational Harmonic Analysis: Special Issue on Wavelet Applications in Engineering, vol. 10, pp. 290-315, May 2001.
- [9] "JPEG2000 part I final draft international standard," ISO/IEC JTC1/SC29/WG1 N1890, Sept 2000.

- [10] Wei Yu, "Integrated rate control and entropy coding for JPEG2000," IEEE Proceedings of the Data Compression Conference, 2004
- [11] C. E. Shannon, "A Mathematical Theory of Communication," Bell Syst. Tech. J., vol.27, pp.379-423 and 623-656, 1948.
- [12] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice Hall, 1971.
- [13] A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation and Compression*. New York, NY: Plenum, 1988.
- [14] Hsueh-Ming Hang and Jiann-Jone Chen, "Source model for transform video coder and its application – Part I: Fundamental theory," IEEE Transaction on Circuit and System Video Technology, vol. 7, NO. 2, April. 1997.
- [15] Antonio Ortega and Kannan Ramchandran, "Rate-distortion methods for image and video compression," IEEE Signal Processing Magazine, pp. 23-50, Nov, 1998
- [16] H. Everett, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources," Operations Research, vol. 11, pp. 399-417, 1963
- [17] ISO/IEC JTC1, "Joint Scalable Video Model (JSVM) 1.0 Reference Encoding Algorithm Description" ISO/IEC JTC1/WG11 N6899, Jan. 2005