BIT RATE CONTROL FOR REAL-TIME MULTIPOINT VIDEO CONFERENCING

by

Xiaoping Hu

BS, Shanghai Jiaotong University, 2001

Submitted to the Graduate Faculty of

School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2003

UNIVERSITY OF PITTSBURGH

SCHOOL OF ENGINEERING

This thesis was presented

by

Xiaoping Hu

It was defended on

July 24, 2003

and approved by

Ching-Chung Li, Professor, Department of Electrical Engineering

Heung-no Lee, Assistant Professor, Department of Electrical Engineering

Thesis Advisor: Luis F. Chaparro, Associate Professor, Department of Electrical Engineering

ABSTRACT

BIT RATE CONTROL FOR REAL-TIME MULTIPOINT VIDEO CONFERENCING

Xiaoping Hu, MS

University of Pittsburgh, 2003

With the rapid development of video compression and network technology, real-time video communications has become a popular part of our daily life. Rate control is needed to satisfy the expectation of high quality and to make it possible to transmit over limited bandwidth. The objective of this thesis is to design a rate control scheme for a real-time Transcoding-Compositing Multipoint Video Conferencing System, which operates exclusively in the DCT domain.

In this Transcoding-Compositing system, the mode of the composited frame should firstly be decided before encoding the composited image. A mode decision method relying on Karhunen-Loeve scene change detection is proposed.

A new linear source Rate-Distortion model is developed in the ρ- domain (ρ is the percentage of zero), based on which rate control scheme is designed. The designed rate control scheme is parted into three levels: Frame Level, Sub-frame Level, and Macroblock Level. Frame Level rate control decides the bit budget for each frame based on the buffer fullness. Sub-frame Level rate control optimizes the distribution of the bit budget among the decimated sub-images. Based on the linear source model, Macroblock Level rate control carries out an adaptive procedure to precisely control the number of encoding bits for each sub-image.

ACKNOWLEDGMENTS

I would like to thank Dr. Chaparro for his support and assistance in writing this thesis. I would like to thank Dr. Li for his help and suggestions. I would also like to thank Jian Xu, James Chen and Haki Ilgin for their help. Finally, let me give my gratitude to Pittsburgh Digital Greenhouse for its partial financial support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 Background And Motivation

With the rapid development of video compression and network technology, real-time video communications has become part of our daily life. To satisfy the expectation of high quality and to make it possible to transmit over a limited bandwidth, several standards have been established to support different services. H.263 succeeds in video conferencing. MPEG-2 and MPEG-4 have been applied to multimedia applications. In these standards, Discrete Cosine Transform (DCT), Motion Estimation, Motion Compensation, Quantization, and Variable Length Coding (VLC) are used to reduce the large volume of digital video data to an acceptable level.

To achieve smooth transmission and to avoid loss of data, the encoded bits are placed into a buffer before being sent. And a decoder buffer is also used at the receiver end as shown in Figure 1-1. A target bit number

$$B_T = \frac{C}{F} \qquad (1\text{-}1),$$

is allocated to each frame, where C is the channel rate and F is the frame rate. When one or more frames occupy more than B_T bits, extra bits are accumulated in the buffer and buffer delay is increased (in the case of real-time video conferencing, the end-to-end buffer delay needs to be very small). Once there are a large number of extra bits in the buffer, the encoder skips one or more frames to reduce the buffer delay and avoid buffer overflow. Skipping frames may cause

undesirable video discontinuity appearing as jitters. On the other hand, when there are frames occupying less than B_T bits, the buffer is underflow and bandwidth is wasted.

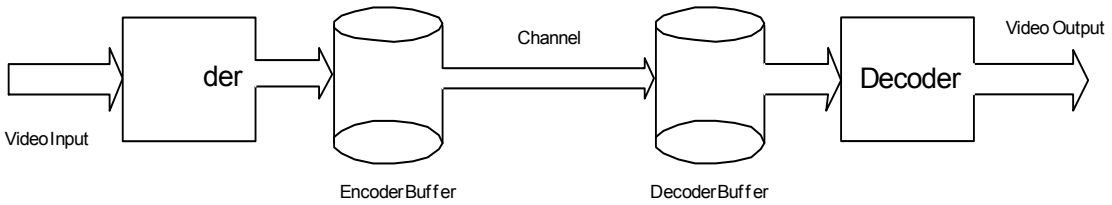

Figure 1-1 Video Communication process

To avoid frame skipping and buffer underflow, an accurate rate control scheme is needed to keep the frame bit count to follow precisely the target bit count B_T. A widely used approach is the operational rate-distortion based approach [1]. An *R-Q* curve[1] is modeled with several control parameters, estimated from coding statistics. A proper quantization step is then picked according to the modeled curve and the target bit. It is obvious that better video quality is achieved when more bits are allocated to a certain frame. Therefore, some bit allocation algorithm should be included in the rate control to guarantee a desirable image quality.

A lot of effort has been put into the rate control issue in recent years. Various rate control tools have been adopted in test models. In TM5[2], the problem is solved in three steps [2]: (1) To allocate bits to each picture according to image activities and buffer fullness; (2) to set nominal slice quantization parameters[3] to meet the target rate; and (3) to derive quantization step of each Macroblock from its nominal quantization parameters according to its activities. The *R-Q* model

[1] *R-Q* curve is a curve which shows the relationship between Bit Rate and Quantization.
[2] TM5 is Test Model 5 for MPEG-2 video stream.
[3] Nominal slice quantization parameter is a coarse quantization step set for the whole frame.

in TM5 is a complex one with three parameters to be decided. TMN8 is based on a quadratic *R-Q* model. Rate control is done on the Frame Layer and Macro Block Layer respectively. Lagrange optimization method is used to calculate the bit allocation on each frame and macro block to achieve good image quality. Though much work has been done, rate control remains an open issue in video communications. No standards have been established concerning rate control, so each encoder can choose its own rate control method freely.

All in all, the objective of rate control is to achieve a certain target bit rate with consistent and desirable visual quality. Based on a new linear source model, a new rate control scheme is proposed in this thesis for a real-time multipoint video conferencing transcoder, which operates in the DCT domain. Given the accurate source model and the moderate computing complexity, the proposed method is very suitable for real-time multipoint video conferencing. Experiments show that it works well at low bit rate communication.

1.2 DCT-Transcoding And Compositing System For Multipoint Video Conferencing

Customers taking part in a videoconference send their video streams to the "bridge" or multipoint control unit (MCU) for decompression, compositing and compression. The composited image, with various required configurations, is sent back to each customer. In real-time multipoint video conferencing, based on the H.263 standard, latency caused by decompressing and recompressing can be reduced and the composited image quality can be improved by using DCT domain exclusively. The block diagram of DCT-based Transcoding and Compositing system is shown in Figure 1-2[1].

[1] Please refer to PDG final report [3] for details.

In this system, computationally expensive components — DCT and IDCT are avoided. The traditional decoder is replaced by a transcoder that transforms H.263 decoder input data into JPEG type of arrays in the DCT domain. Motion vectors of the incoming streams are used to calculate the motion vectors of the composited stream with extremely low computing complexity. Experiments have shown that in processing an inter frame, the speed up of DCT-based procedure versus the conventional space domain procedure is almost 20 times. Figure 1-3 shows the comparison of speed of them.

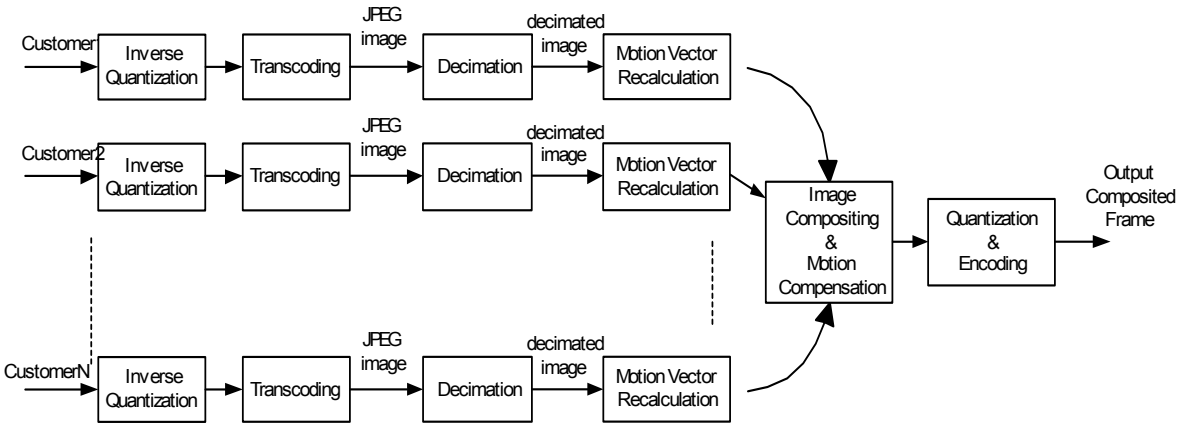

Figure 1-2 DCT Transcoding-Compositing Multipoint Videoconference System (refer to PDG final report [2] for details)

In the compressed domain processing, decimation is done in DCT domain. In the space domain, decimation is implemented by simply deleting rows/columns or by averaging related pixels. Error may occur because of aliasing. The DCT domain decimation improves the quality of the composited image by avoiding aliasing.

Discrete Cosine Transform, as an orthogonal linear transform, guarantees the same standard deviation, a crucial parameter indicating the complexity of an image, before and after

transform. Although the DCT images lack visualization, the information it contains provides important clues of the image property for rate control and other processing.

The proposed rate control scheme is based on the DCT domain transcoding and compositing system. The system model is stated in detail in the next section.

1.3 Statement Of The Problem

Assume that during a videoconference, four sources are sending video streams as shown in Figure 1-4. The received images are decimated and composited in real-time by the recommended DCT Transcoding-Compositing system. Some rate control method is to be designed to make sure that each customer gets consistent quality composited frames over a certain channel bit rate. Some assumptions are:

(1) Communication between customers and transcoder are based on H.263 low bit rate video codec standard.

(2) The incoming frames are of good quality, rate controlled images.

(3) The incoming frames have the same frame rate, or there already exists some frame skipping/interpolation component insuring that all incoming frame are synchronized.

(4) GOP[1] structures are known to the transcoder and remain unchanged during the communication. All customers have the same form of GOP structure IPP…PPP, that is, each GOP contains one intra frame followed by a number of inter frames. The number of inter-frames in a GOP may be variant. Therefore, it is possible that the transcoder receives frames of different mode at some epoch.

[1] GOP stands for group of pictures.

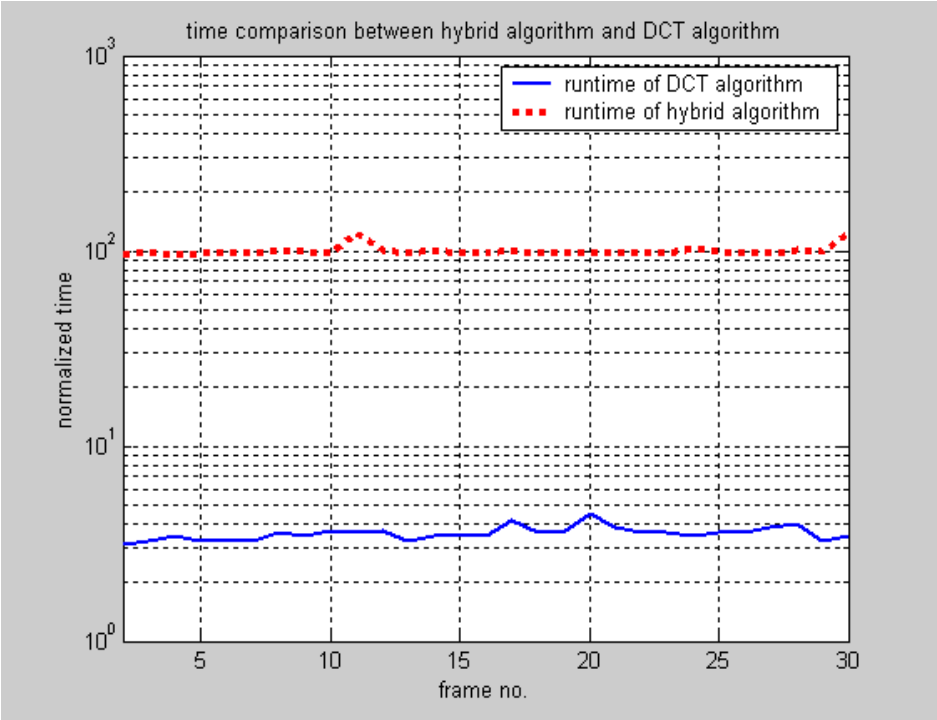

Figure 1-3 Time Comparison between Hybrid Algorithm and DCT Algorithm

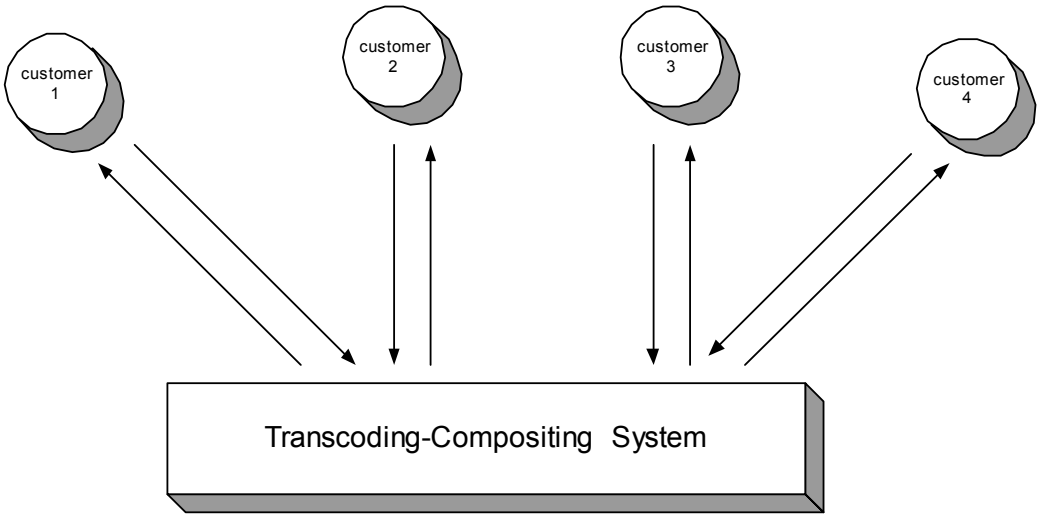

Figure 1-4 Multipoint Video Conferencing

(5) Quantization is macroblock based.

(6) Macroblock based mode, that is, every macroblock may be inter or intra.

Based on the system model (Figure 1-2) and the assumptions above, there are three issues to consider when designing a rate control scheme.

(a) Mode Decision and Mode Transform

As discussed in assumption (4), since the number of P frames in a GOP may be different among customers, intra-frames may appear at different epoch in the incoming streams. Therefore, the transcoder should be able to perform mode decision and reform intra- and inter-frame before doing compositing.

When the transcoder decides to change an inter frame into an intra frame, motion vector recalculation is saved and a reconstructed frame is transmitted as if it is an intra frame. Neither quality degrading nor upgrading happens in this case. But a large number of bits are needed. The number of bits needed to code an intra frame is usually 10 to 20 times of that needed to code an inter frame [4].

When changing an intra frame into an inter frame, great quality degrading occurs in case of scene change. To save bits and to guarantee the image quality at the same time, a scene change detector is needed to decide the sub-image mode. If scene change occurs in some incoming stream, the corresponding sub-image is coded as the mode of intra. Otherwise the sub-image is treated as inter mode.

(b) Bit Allocation among Sub-images

Four incoming frames are composed into one frame after decimation. It is undesirable to allocate the same bit count to each sub-image. Images with a lot of motion or a high level of complexity should be assigned more bits to achieve certain quality, while other images do not

need as many bits. A bit allocation algorithm is needed to cleverly assign bits in order to optimize the overall quality of the sub-images.

(c) Source Distortion Model and Rate Control Algorithm

Rate control is achieved by selecting a proper quantization step according to a source distortion curve. The complexity and accuracy of the source model are two concerns in real-time communication. Usually an accurate model cannot be derived without pre-encoding the source image several times [2]. This is not possible in real-time conferencing since it is highly time consuming. In this thesis, a rate control algorithm, based on a new simple source model, which encodes every macroblock once, is proposed based on a new simple source model.

Mode decision, bit allocation and new source model are discussed in detail and some creative rate control algorithms are proposed to solve these problems in later chapters.

1.4 Outline

This thesis is comprised of 6 chapters. The present chapter has introduced the motivation, background and research objective for real-time multipoint video conferencing rate control. Chapter 2 deals with the problem of mode decision. A scene change detection method based Karhunen-Loeve (KL) feature extraction is proposed. Chapter 3 reviews some existing rate distortion models. A new linear source model is introduced and the justification for this model is given. Based on this new model, rate control and bit allocation algorithms are developed on the level of frame, sub-frame and macroblock respectively in Chapter 4. Experimental results are displayed in Chapter 5. Finally, in Chapter 6, discussions and conclusions are given, as well as the issues to be explored further.

2.0 MODE DECISION METHOD

2.1 Introduction

In the H.263 standard, prediction is done for inter pictures. The coding mode using prediction is called *inter*, and coding mode using no prediction is called *intra*. Intra coding is signaled at the picture level (I frame for intra or P frame for inter) or at the macro block level in P frames. Details about the mode specification of H.263 are stated in Appendix A. Briefly, if one frame is set to be an I frame, all of its macroblocks are of intra mode and no motion estimation or motion compensation happens. However, macroblocks in an inter frame can be redefined as intra macroblocks in macroblock layer.

In the transcoding-compositing system, the mode of the composited frame should be decided according to the input frames. When all of them are inter/intra frames, the composited frame is simply set to be inter/intra frame. Due to the various properties of the source streams in videoconferencing, it is hard to require that customers send frames in the same mode during the conference.

The simplest way to combine four sub-frames with different mode is to let each sub-frame keep their original mode. It is not an ideal choice, however. Encoding sub-frames in intra mode is expensive in terms of bit count. It has been shown that an intra frame needs up to 20-30 times the numbers of bit of an inter frame [4]. To save bit count and avoid unnecessary intra sub-frames, scene change detection is used to decide whether an intra sub-frame is necessary or not.

When scene change happens, intra sub-frame cannot be saved. If no scene change is detected, intra sub-frame is reformed into inter sub-frame without degrading image quality much.

In the next section, two traditional scene change detection methods are described briefly. And in Section 2.3, a new method based on Karhunen-Loeve (KL) algorithm is introduced. Comparison of these methods is shown. Finally, in Section 2.4, the summary of the overall mode decision algorithm is given.

2.2 Traditional Methods Of Scene Change Detection

Scene change detection is the process of dividing a video stream into shots based on the content of the video, where a shot is a sequence of continuous frames representing a continuous action in time and in space. Generally speaking, there are two types of transitions: abrupt scene change and gradual scene change. An abrupt scene change is a sharp transition, which exists only between two frames, whereas a gradual scene change requires several frames to complete. There has been a lot of work reported on detecting scene change. Usually difference matrices are used to evaluate the changes between successive frames and global thresholds are used to determine whether changes have taken place [5][6][7]. Normally, only luminance is used to detect change.

DC images, consisting of DC values of each block, are spatially reduced versions of the original images. Such spatially reduced images are useful in scene change detection. In the Transcoding-Compositing system, JPEG image in DCT domain is produced by the transcoder. DC matrix can be easily extracted from it.

Conventionally there are two types of measurements of changes between frames: DC pixel-based difference and DC histogram-based difference [5][6].

2.2.1 DC-pixel Scene Change Detection [5]

This method relies on the measurement of corresponding pixel difference between two successive DC matrices. The DC image difference is defined as:

$$d_{DC}(X,Y) = \sum_{n=1}^{N}\sum_{m=1}^{M} | C_X(n,m) - C_Y(n,m) | \qquad (2\text{-}1),$$

where $C_X(n,m)$ and $C_Y(n,m)$ are DC images of frame X and Y, and $N \times M$ is the number of blocks in each frame. Wherever d_{DC} is larger than a predefined global threshold d_{th}, a scene change is detected between frame X and Y.

The greatest advantage of this method is its simplicity. Computing complexity is extremely low in this case. Only $N \times M$ times of subtraction and $N \times M - 1$ times addition are involved. However, pixel based techniques may generate false alarms whenever there are moving objects or camera movement in the video.

2.2.2 DC-histogram Scene Change Detection [6]

The difference based on DC histogram is defined as:

$$d_{hist}(X,Y) = \sum_{i} | h_x(i) - h_y(i) | \qquad (2\text{-}2),$$

where $h_x(i)$, $h_y(i)$ are the *ith* bin in the histogram of DC matrices of image X and Y. Obviously the histograms of the DC matrices of frame X and Y should have the same range in this case.

The author of [6] suggests using another type of histogram difference called χ^2 test difference:

$$d_{\chi^2} = \begin{cases} 0 & \\ \sum_{j=1}^{K} \dfrac{(H_X(j) - H_Y(j))^2}{\max(H_X(j), H_Y(j))} & \text{if } H_X(j) \neq 0 \text{ or if } H_Y(j) \neq 0 \end{cases} \qquad (2\text{-}3),$$

where $H_X(j)$ is the bin value of histogram of frame X and K is overall number of bins.

Histogram based techniques are fairly immune to the effect of motion, but they may miss scene changes if the luminance distribution of the frames does not change significantly. This method does not work well in case of gradual scene change.

2.3 Scene Change Detection Based On KL Expansion

Both DC pixel-based detection and histogram-based detection tend to give false detection or miss scene change. A novel method based on the Karhunen-Loeve (KL) expansion is proposed in this section, which provides better detection with moderate computing complexity.

Consider the images in one frame shot as in the same data class and the images subject to scene change in different data classes. The task of scene change detection is to detect whether two succeeding images have different features and belong to different classes. For an image, the original measurement of its feature is assumed to be the DC value matrix in this case. The Karhunen-Loeve (KL) Expansion is a very efficient way to reduce a high dimensional space to a low-dimensional subspace and capture the feature of a data class effectively. The vectors produced by the Karhunen-Loeve projection are typically called the principal components.

As an effective method of feature extraction, the KL expansion has been widely used in the field of Pattern Recognition. To the best of our knowledge, it is the first time to be applied in scene change detection.

2.3.1 Karhunen-Loeve (KL) Expansion

The images of *m*-row and *n*-column DC-value matrix can be represented by $N = m \times n$ dimensional vector. Typically, N is large. For CIF images decimated by 2, which have $\frac{288}{2} \times \frac{352}{2}$ pixels, DC matrix is of the size 18×22, so $N = 396$ in this case. Here N is still very large, but you can see further reduction of N in the next section.

Let $X \in R^N$, be a random vector representing an image, where N is the dimensionality of the DC image space. Concatenating the rows or the columns of the DC image forms the vector. The autocorrelation matrix of X is defined as:

$$C_X = XX^T \qquad \text{(2-4)},$$

where T denotes the transpose operation, and $C \in R^{N \times N}$. We can factorize the autocorrelation matrix C_X into the following form:

$$C_X = \Phi \Lambda \Phi^T \qquad \text{(2-5)},$$

with $\Phi = \{\phi_1, \phi_2, ... \phi_N\}$, and $\Lambda = \{\lambda_1, \lambda_2 .... \lambda_N\}$, where $\Phi \in R^{N \times N}$ is an orthonormal eigenvector matrix and $\Lambda \in R^{N \times N}$ a diagonal eigenvalue matrix with diagonal elements in decreasing order $(\lambda_1 \geq \lambda_2 \geq ... \lambda_N)$, and ϕ_1, ϕ_2, ... ϕ_N and $\lambda_1, \lambda_2 .... \lambda_N$ are the eigenvectors and the eigenvalues

of C_X, respectively. In the Karhunen-Loeve projection only a subset of principal components is used for the transformation matrix construction $P = \{\phi_1, \phi_2, ...\phi_m\}, m < N$.

The lower dimensional vector Y captures the most expressive features (MEF) of the original data X,

$$Y = P^T X \qquad \text{(2-6).}$$

Y is the best dimension reduced representation of X in the sense of mean-square-error. Consider a reconstructed X using KL transform matrix P,

$$\overline{X} = PY \qquad \text{(2-7)}$$

The mean square error is defined as

$$Err = (X - \overline{X})(X - \overline{X})^T \qquad \text{(2-8).}$$

[8] shows that for a given m, that is , for a finite number of basis functions used in the expansion, the KL expansion minimizes the mean square error. So KL expansion provides the best representation of the original DC image with least loss of information.

2.3.2 Scene Change Detection Using KL Expansion

The algorithm of scene change detection using KL expansion is shown in Figure 2-1. Based on the KL Expansion, the difference of two succeeding frames is defined as

$$d_{KL}(J1, J2) = \sum_{i=1}^{m} | Y_{J1}(i) - Y_{J2}(i) | \qquad \text{(2-9),}$$

where $Y_J(i)$ is the *ith* component of reduced feature vector Y of image J.

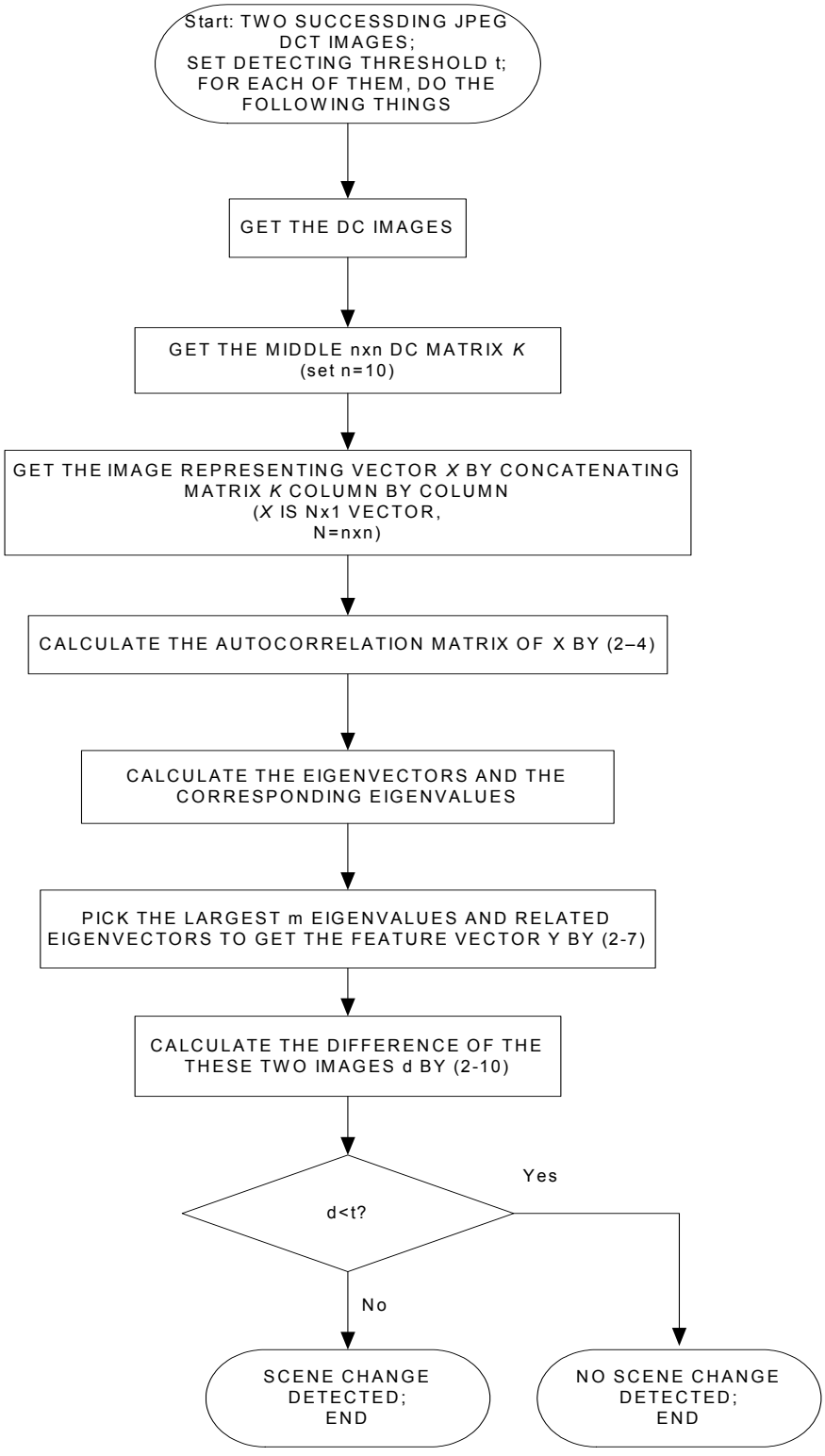

Figure 2-1 KL Algorithm

Most of videoconference frames are head-shoulder images, on which the most important part focuses on the middle part of the image. To reduce the computing time, only the middle part the DC image is used to extract the image feature. For CIF images decimated by 2, whose DC image is 18×22, the central 10×10 DC values are adequate to successfully extract the image feature. The dimension of the reduced feature vector can be very small. Experiments show that m=3 can produce desirable results. In Figure 2-1, threshold t is decided from experiments off-line or can be learned on-line.

2.3.3 Experimental Results

To test the performance of the scene change detection method, a video of 600 frames long was collected. It consisted of 8 concatenated sequences, such that there are 7 abrupt scene changes occurring between two succeeding sequences. And there is one gradual scene change in the sequence "Coastguard". The list of the sequences is shown in Table 2-1.

The differences calculated based on DC_pixel method, DC_histogram method and KL method are shown respectively in Figure 2-2. For KL method, the central 10x10 DC image was selected as the representative region and m=3. For each of the three plots, there are peaks at the seven abrupt scene changes. In case of DC_pixel difference, there is a small peak at about 410^{th} frame, representing the gradual scene change there. Fast motion causes large DC_pixel difference (see the 2^{nd} and 3^{rd} sequences). And abrupt scene change is not obvious enough sometimes (see the fourth peak). Therefore, it is hard to decide a global threshold for DC_pixel difference. High missing detection rate may occur with a large threshold and high false detection rate may occur with a small threshold. Gradually changing scene would not be detected in many

instances. In case of DC_histogram difference, fast motion does not affect much. Gradual scene change almost disappears. In the third case, with the KL method, all the scene changes are reflected obviously by the difference. Although fast motions produce relatively large differences, a global threshold can still be easily set to detect the scene changes.

The performance comparison of these three scene change detection algorithms is demonstrated in Table 2-2. There are 100 scene changes among the 6000 tested frames, of which 10 are gradual scene changes. The simplest DC_pixel scene change detection has a missing detection rate of 18%, and false detection rate 9%. The DC_histogram scene change detection produces a missing rate 10% and false detection rate 2%. And with the KL method, missing rate reduces to 3%, and false detection rate 2%. The KL method outperforms the conventional scene change detection methods by more than 5%.

Table 2-1 Generated Video Sequences

Sequence Name	Frame Number and Type
Miss America	1~90 (intraframes occur at 1st, 31st and 61st frame, no scene change in between)
Football	91~150 (intraframes occur at 91st and 121st frame, fast motion occurs, no scene change in between)
Flower	151~210 (intraframes occur at 151st and 181st frame, fast motion, no scene change in between)
Claire	211~270 (intraframes occur at 211th and 241st frame, no scene change in between)
Trev	271~360 (intraframes occur at 271st, 301st, and 331st frame, no scene change in between)
Coastguard	361~480 (intraframes occur at 361st, 391st, 421st and 451st frame, gradual scene change occurs between 410th ~420th frames)
Foreman	481~540 (intraframes occur at 481st and 511th frame, no scene change in between)
Salesman	541~600 (intraframes occur at 541st and 571st frame, no scene change in between)

Figure 2-3 demonstrates the computation time spent on each of the three methods. We can see the time consumed on the KL based scene change detection is between that on the DC-pixel detection and that on the DC-histogram detection. The KL-based scene change detection can achieve a good performance while having reasonable computational complexity.

2.4 Summary Of Mode Decision Algorithm

In the transcoding-compositing system, the transcoder converts all incoming inter/intra frames into intra-frames. We also have the information of how the GOP of the input streams was organized. With these two pieces of information we can use the scene detection to keep or change the known GOP structure of the input streams.

When the four incoming frames are of different modes, mode decision algorithm is applied on the intra sub-frames to achieve both low bit rate and desirable video quality. Scene change detection techniques discussed above are used to decide whether it is necessary to use intra mode sub-frame or not. If there is no scene change, the incoming intra sub-frame is subjected to motion estimation as an inter sub-frame. If there is scene change detected, intra mode is set for the sub-frame.

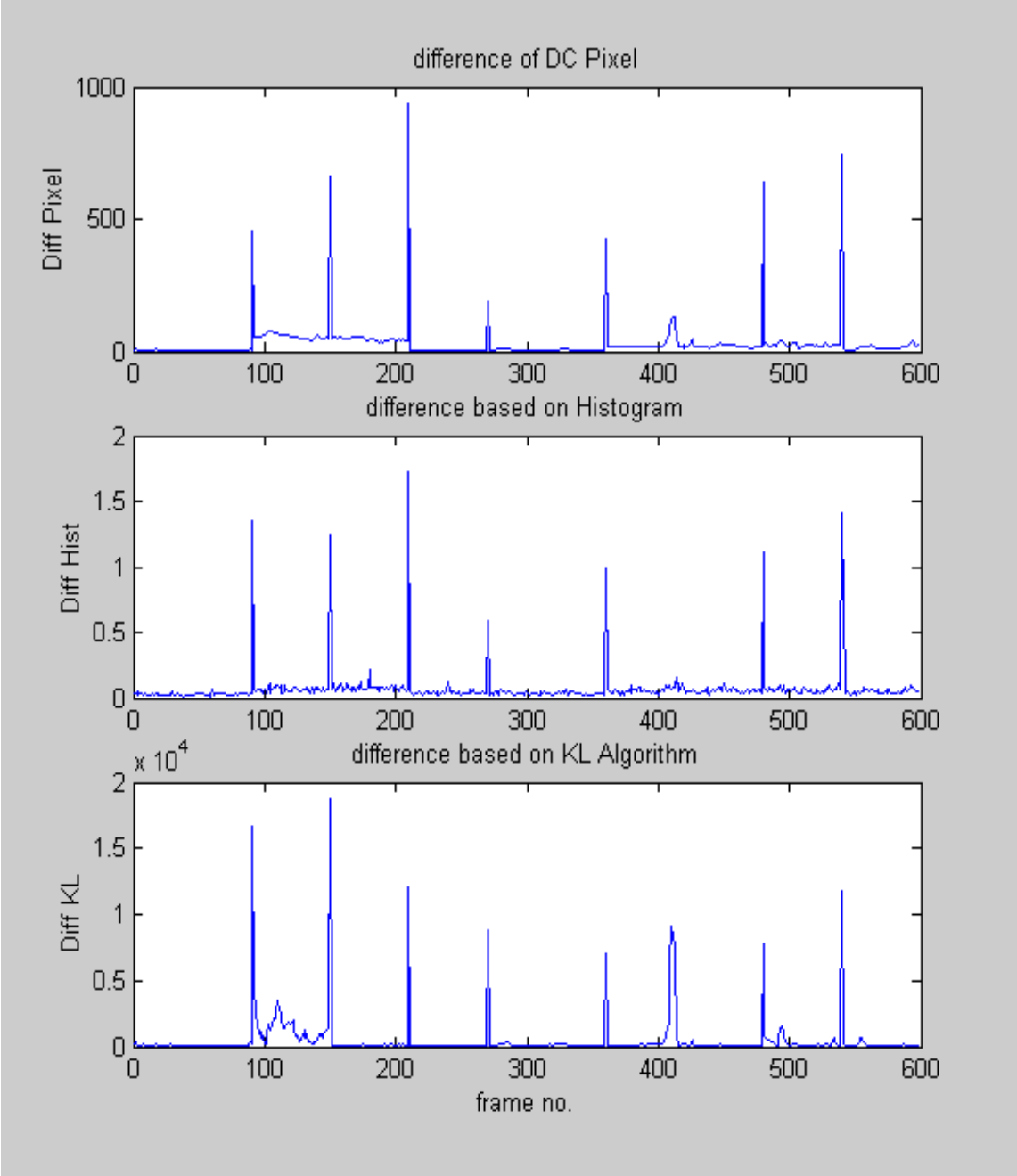

Figure 2-2 Comparison of the Resulting Difference Based on three methods

Table 2-2 Performance Comparison of Scene Change Detection Algorithms

Method	Number of Frame Tested	Number of Scene Changes	Missing Detection	False Detection
DC-pixel	6000	100	18	9
DC-histogram	6000	100	10	2
DC-KL (central 10x10 DC image is used, m=3)	6000	100	3	2

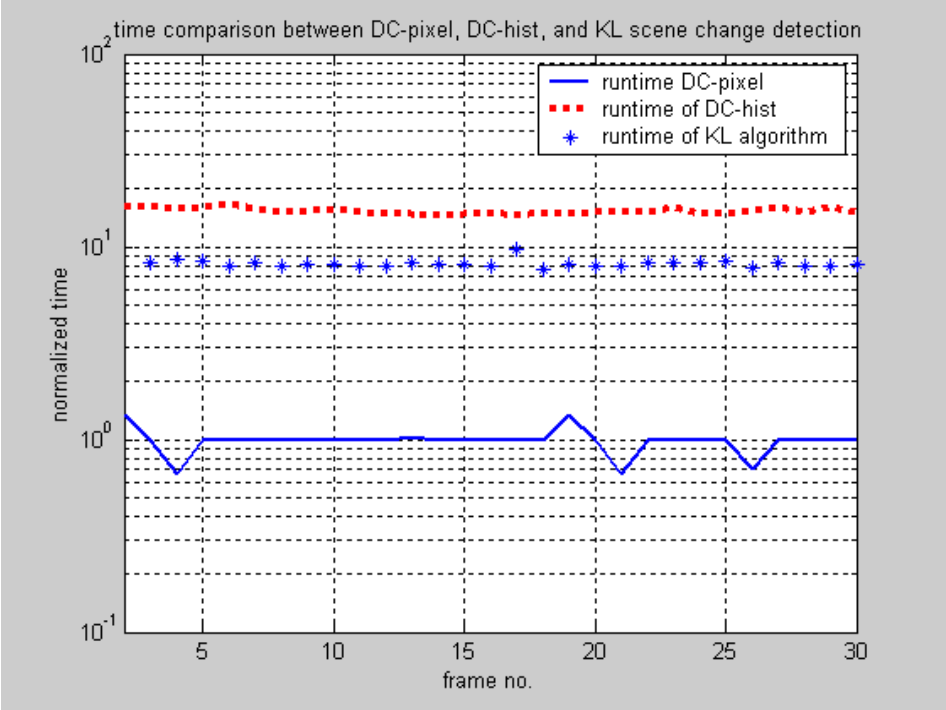

Figure 2-3 Time Comparison between DC_pixel, DC_hist, and KL Scene Change Detection Algorithm

3.0 RATE-DISTORTION MODEL

3.1 Introduction

In standard video coding, the target bit rate R_T (average bit number for each coefficient), is controlled by the quantization. In H.263, uniform quantization is performed before Run Length Coding. The relationship between the quantization step and the bit rate is described by rate-quantization (*R*-*Q*) curve, denoted by $R(Q)$. Almost all the existing rate control methods rely on the *R*-*Q* curves by selecting the corresponding quantization step $R^{-1}(R_T)$.

This section focuses on how to derive an accurate source rate distortion model. In Section 3.2, the classical rate distortion theory is given. In Section 3.3, we review briefly the existing rate control algorithms. Then the linear relationship between the coding bit rate R and the percentage of zeros among the quantized coefficients, denoted by ρ, is shown in Section 3.4. The justification of this relationship is given and experiments show the accuracy of this model.

3.2 Classical R-D Model Based On Rate Distortion Theory

For a communication system shown in Figure 3-1, the rate distortion theory calculates the minimum transmission bit rate R for a required picture quality. Results of rate distortion theory are obtained without consideration of a specific coding method.

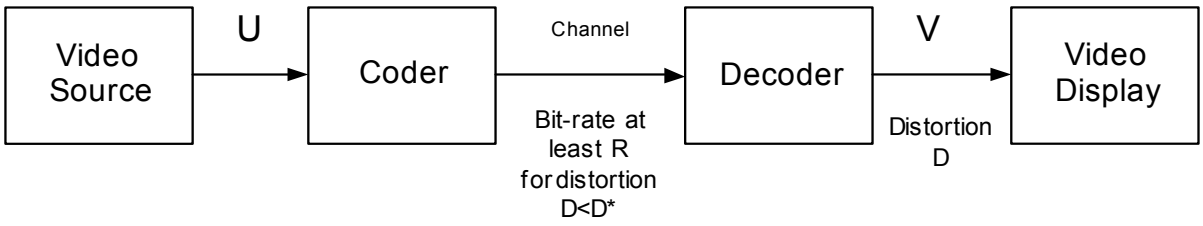

Figure 3-1 Communication System with Distortion

The Rate Distortion Function is defined as [9]:

$$R(D^*) = \min_{D \le D*} \{I(U,V)\} \qquad (3\text{-}1),$$

where $I(U,V)$ is the average mutual information of symbols U and V, D is the average distortion under some distortion measurement. Equation (3-1) shows that for a given maximum average distortion D^*, the rate distortion function $R(D^*)$ is the lower bound for the transmission bit-rate.

Under the ideal condition that the source coder produces distortions U-V that are statistically independent from the reconstructed signal V, $R(D^*)$ can be estimated by Shannon Low Bound as

$$R(D^*) \ge H(U) - \max_{D \le D*} H(U - V) \qquad (3\text{-}2),$$

where $H(U)$ is the entropy of symbols U.

Equation (3-2) is used to estimate the distortion rate function for various information sources. An image source is usually considered to be a Laplacian source. A Laplacian distribution is given by:

$$p_l(x) = \frac{\lambda}{2} e^{-\lambda|x|} \qquad (3\text{-}3),$$

where λ is a parameter depended on the variance of the source. Figure 3-2 shows the distribution of the DCT coefficients of an inter frame in the sequence Miss America, and the corresponding Laplacian model. The Laplacian model matches the image source very well.

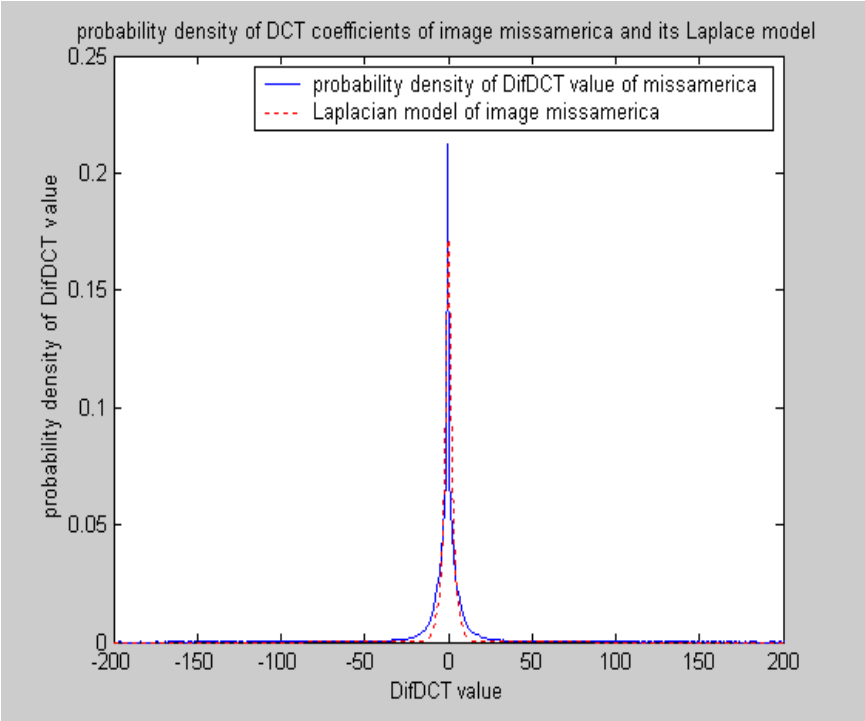

Figure 3-2 Probability Density of DCT Errors of Image Miss America and its Laplacian Model

According to [9], for Laplacian sources, defining the distortion measure as

$$D(x,\bar{x}) = |x - \bar{x}| \qquad (3\text{-}4),$$

where x is the input symbol and $\bar{x}$ is the output symbol of the encoder, the lower bound of the bit rate is

$$R(D) = \log_2\left(\frac{1}{\lambda D}\right) \qquad (3\text{-}5)$$

3.3 Existing R-Q Models

The R-D model for image source based on rate-distortion theory, developed in the previous section, is unfortunately not an accurate model. In the classical R-D formula, the only parameter describing the image source is λ, which depends on the variance of the image source. Various coding algorithms are not taken into consideration. The reality is, even for the same image, different coding algorithms will lead to quite different R-D functions.

Since the classical R-D model does not work well in practical coding applications, operational approaches modeling the relationship of Rate and Quantization have been developed. An exponential model is proposed in [2]. The model parameters are estimated from the coding statistics generated by re-encoding of the input video. This method is not suitable for real-time application because of its high computational complexity.

In [10], a second order quadratic R-Q model is developed where coding statistics of previous frames are utilized to estimate the parameters of the current frame. Computing time is reduced but the algorithm does not work well when two succeeding frames have different parameters.

In TMN8, a quadratic R-Q model is proposed with an adaptive quantization scheme at the MB level. This method has a superior rate control performance compared with previous models [1]. The image source is modeled as:

$$R \approx K \frac{\sigma^2}{Q^2} \qquad \text{(3-6)},$$

where σ^2 is the variance of the source data, Q is the quantization step size, and K is the parameter to be estimated during the encoding procedure.

3.4 Linear Model In ρ Domain

All of the R-D models discussed in previous sections attempt to find a formula relating the coding bit rate R and distortion D with the quantization step Q. Almost all the previous models are carried out in the Q – domain.

All the R-D analysis carried out in Q – domain are quiet rough, because in reality Rate and Quantization have a complex non-linear relation that is hard to represent with a simple expression with a few parameters.

The inaccuracy is enlarged in our case. In the Transcoding-Compositing system presented in Chapter 1, quantized frames from several customers are dequantized in the transcoder before further processing. And re-quantization is performed on the composited image to achieve certain bit rate. The image subjected to quantization and dequantization no longer matches the R-Q model very well.

In [11], the author observed that in transform coding of images and videos, the percentage of zeros in the quantized transform coefficients, denoted by ρ, has a critical effect on the coding bit rate. When the distribution of the transform coefficients is continuous and positive, ρ monotonically increases with the quantization step Q. This implies that there is a one-to-one mapping between Q and ρ. Hence, the relation between bit rate R and distortion D are also functions of ρ. A study of the R-D model as function of ρ is termed ρ – domain R-D analysis.

In the literature [11], [12], [13], [14] and [15], it has been shown that a simple linear relation between ρ and coding bit rate for standard video sources is developed. In this section, it is demonstrated that this linear source model also works on our Transcoding-Compositing

system, if the re-quantization step is much larger than the quantization step performed on the incoming frames.

In Section 3.4.1, the linear source mode in ρ – domain is proposed and experiment results show that it works well on our system. In Section 3.4.2, a theoretical justification is provided for this linear model. The physical meaning of the model parameter is also discussed.

3.4.1 Proposed Source Model

Simulations in [11] show that in all typical transform coding systems, such as the wavelet-based image compression, JPEG image coding, MPEG-2, H.263, and MPEG-4 video coding, $R(\rho)$ is approximately a linear function:

$$R(\rho) = \theta \cdot (1 - \rho) \qquad (3\text{-}7),$$

where ρ is the percentage of zeros among the quantized transform coefficients, R is the bit rate of the coded data, θ is a constant directly related to the image content. This leads to a unified source model for all types of transform coding systems. In a rate control scheme, adaptive algorithms only need to estimate the value of θ. Once θ is determined, the rate curve can be constructed according to (3-7).

The linear model also works well on our system provided that the re-quantization step is much larger than the quantization step of the incoming frames. It is satisfied in most of the case in our system. Four sample images randomly selected are shown in Figure 3-3. All these sample images are incoming frames quantized by *QP=5*, that is, quantization step *Q=10*. Dequantization is performed by the transcoder and the resulting JPEG images are decimated by 2, and composited as shown. Each of the decimated images are subjected to H.263 encoding with a

$QP' > 5$. For each QP', we compute the corresponding percentage of zeros, ρ, and record the coding bit rate *R*. By varying QP', we generate a series of points on the rate curve $R(\rho)$, which are plotted in Figure 3-4. It can be seen that $R(\rho)$ is almost a straight line in both case of intra frame and inter frame. In addition, this line passes through the point [1.0,0.0]. This is because, when ρ is 1.0, all the coefficients are quantized to zeros and the corresponding coding bit rate should become zero.

Figure 3-3 Sampled Images

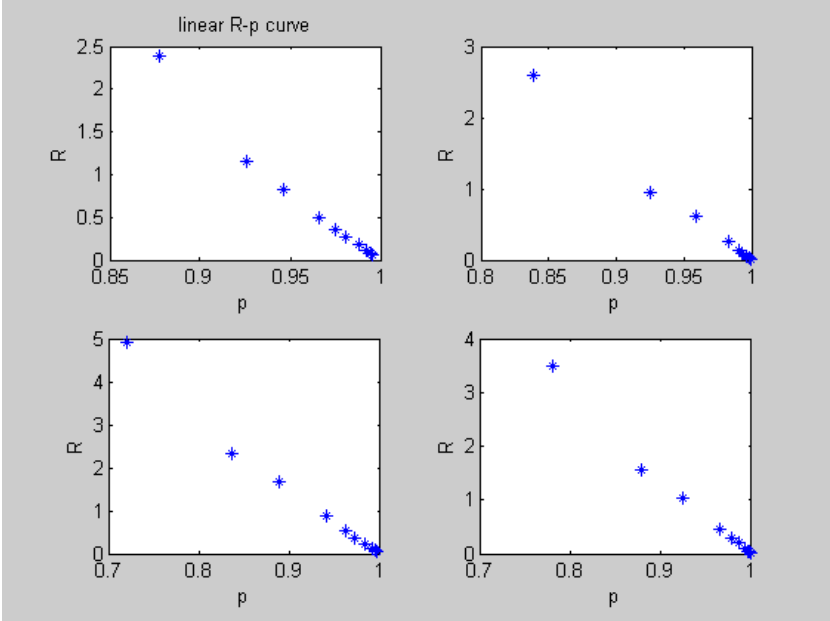

Figure 3-4 Linear $R-\rho$ Curves

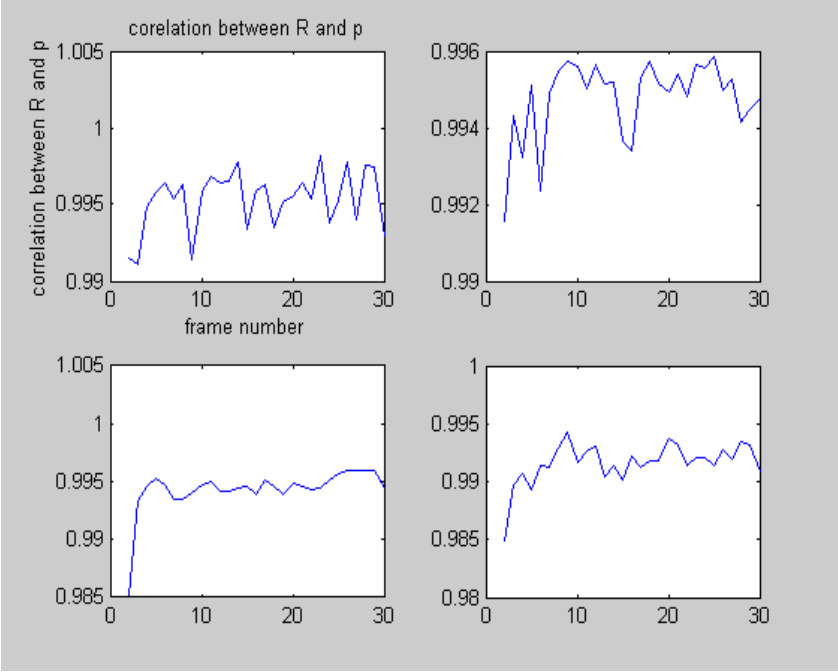

Figure 3-5 Correlation between *R* and ρ

To demonstrate this linear relationship more clearly, the correlation coefficient between R and ρ is given in Figure 3-5. From Figure 3-5, it can be seen that the correlation coefficients are always larger than 0.985, which is extremely close to 1. It implies the linear relationship between R and ρ.

As discussed previously, there is a one-to-one projection between Q and ρ. Therefore, any function in the Q – domain can be mapped to the ρ - domain and vice versa. This mapping can be derived by calculating the distribution of the transform coefficients. This will be discussed further in the rate control scheme shown in Chapter 4. In rate control algorithms, this mapping is stored as a look-up table. The mapping of the rate curve between Q and ρ domains is performed by table look-up.

This proposed model is conceptually simple and has low computational complexity. And it outperforms other rate control models in our case by providing more accurate and robust rate control.

3.4.2 Linear Source Model Justification

Rate Distortion Theory is developed based on Shannon's source coding theorem [9]. The above linear model also can be explained theoretically by this theorem. The model justification is given in this section. This model justification is uniquely developed according to H.263 quantization scheme.

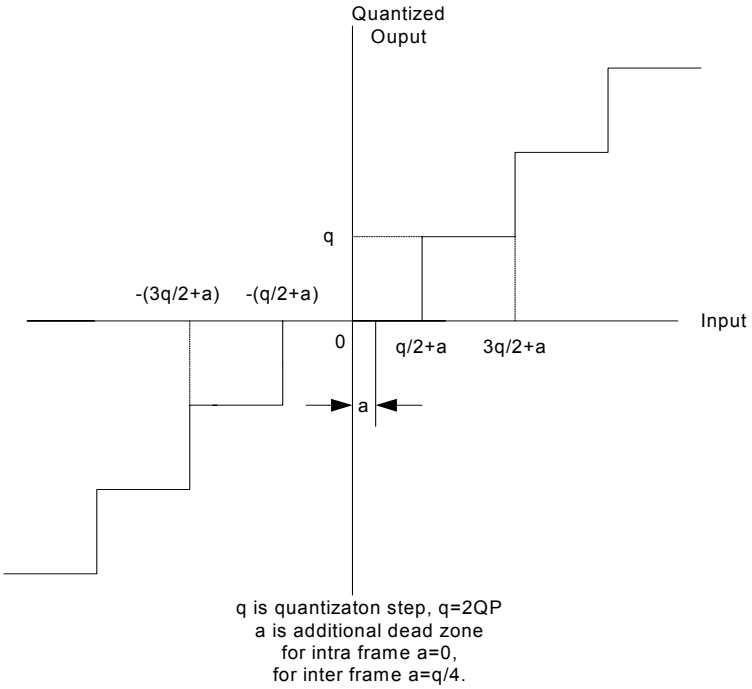

Figure 3-6 Quantization in H.263

The quantization in H.263 is a uniform quantization with a dead zoon around zero[16]. Figure 3-6 shows the quantization rule. For any DCT coefficient x, its quantized value is given by

$$Qx = \begin{cases} round(\frac{x}{8}) \\ round(\frac{x}{Q}) \\ round(\frac{x-\alpha}{Q}) \end{cases} \qquad (3\text{-}8),$$

where the first equation applies for DC coefficients in intra-frames, the second applies for AC coefficients in Intra-frames, and the third applies for the inter-frame coefficients. And Q is the

quantization step and α is an additional dead zone for inter frame coefficients of the quantizer, which is assumed to be $\frac{Q}{4}$ in H.263.

The distribution of the source image and the distortion measurement is given by (3-3) and (3-4). When the re-quantization step is much larger than the quantization step performed on the incoming frames, the quantization-dequantization procedure before decimation can be ignored. The dequantized and decimated image can still be approximately as Laplacian distributed by (3-3). For intra frames, ignoring the DC coefficients, the Distortion can be approximated by

$$D(Q) \approx \int_{-0.5Q}^{0.5Q} p_l(x) x dx + \sum_{\substack{i=-\infty \\ i\neq 0}}^{\infty} \int_{(i-0.5)Q}^{(i+0.5)Q} p_l(x) \,|\, x - iQ \,|\, dx \qquad (3\text{-}9)$$

And the percentage of zero ρ is given by

$$\rho = \int_{-0.5Q}^{0.5Q} p_l(x) dx = \int_{-0.5Q}^{0.5Q} \frac{\lambda}{2} e^{-\lambda|x|} dx = 2\int_{0}^{0.5Q} \frac{\lambda}{2} e^{-\lambda x} dx = 1 - e^{-0.5\lambda Q} \qquad (3\text{-}10).$$

According to Appendix B, (3-9) is equal to

$$D(Q) = \frac{1}{\lambda}[1 + \frac{e^{-\lambda Q}}{1 - e^{-\lambda Q}}(2 - e^{-0.5\lambda Q} - e^{0.5\lambda Q}) - e^{-0.5\lambda Q}] \qquad (3\text{-}11)$$

Combining (3-10) and (3-11), distortion in terms of ρ becomes

$$D(\rho) = \frac{1}{\lambda} \cdot \frac{1 - (1-\rho)}{1 + (1-\rho)} \qquad (3\text{-}12).$$

According to (3-5), the rate function in ρ domain is

$$R(\rho) = \log_2 \left[\frac{1 + (1-\rho)}{1 - (1-\rho)} \right] \qquad (3\text{-}13).$$

Since 1- ρ is always a small value (less than 0.3) in practice, (3-13) can be approximated by Taylor expansion, which yields

$$R(\rho) = K \cdot (1-\rho) + O([1-\rho]^3), \quad K = 2\log_2 e = 2.8854 \qquad (3\text{-}14).$$

Therefore, theoretically, $R(\rho)$ is an approximately linear function.

In case of quantizing coefficients in inter frames, a larger dead zone is applied. The distortion is

$$D(Q) \approx 2\int_0^{0.5Q+\alpha} p_l(x)xdx + 2\sum_{i=1}^{\infty}\int_{(i-0.5)Q+\alpha}^{(i+0.5)Q+\alpha} p_l(x)\,|\,x-iQ\,|\,dx \qquad (3\text{-}15)$$

And the percentage of zeros becomes

$$\rho = \int_{-0.5Q-\alpha}^{0.5Q+\alpha} p_l(x)dx = 1-e^{-(0.5Q+\alpha)\lambda} \qquad (3\text{-}16).$$

After the calculation in Appendix B, the expression of $D(\rho)$ becomes

$$D(\rho) = \frac{1}{\lambda}\left(\frac{1-2t+t^{2a}-2t(a-1)\log(t)}{1-t^{2a}}\right) \qquad (3\text{-}17)$$

and

$$R(\rho) = \log_2\left(\frac{1-t^{2a}}{1-2t+t^{2a}-2t(a-1)\log(t)}\right) \qquad (3\text{-}18)$$

where

$$t = 1-\rho, \quad \alpha = 0.25Q \text{ ,and } a = \frac{0.5Q}{0.5Q+\alpha} = \frac{2}{3}, \qquad (3\text{-}19).$$

The plot of (3-18) is shown in Figure 3-7. It can be seen that the plots are very close to being a straight line. This justifies the linear rate model given by (3-7) for Laplacian image source. Please note that the above justification is based on the assumption that the image DCT coefficients are strictly Laplacian distributed and on the lower bound bit rate equation (3-5). The resulting K in (3-14) cannot be used as θ in reality. The estimation of parameter θ will be introduced in the next chapter.

3.4.3 Physical Meaning of θ

The only parameter of the proposed source model is the slope θ. Obviously, θ is related to some characteristic of the input source image. Intuitively, the slope should become steeper when the complexity of the image goes up, since more bits are needed to code more complex images. This suggests that the value of θ is closely related to the amount of texture presented in the source image. This explains the physical meaning of θ which is the only parameter in the proposed source model.

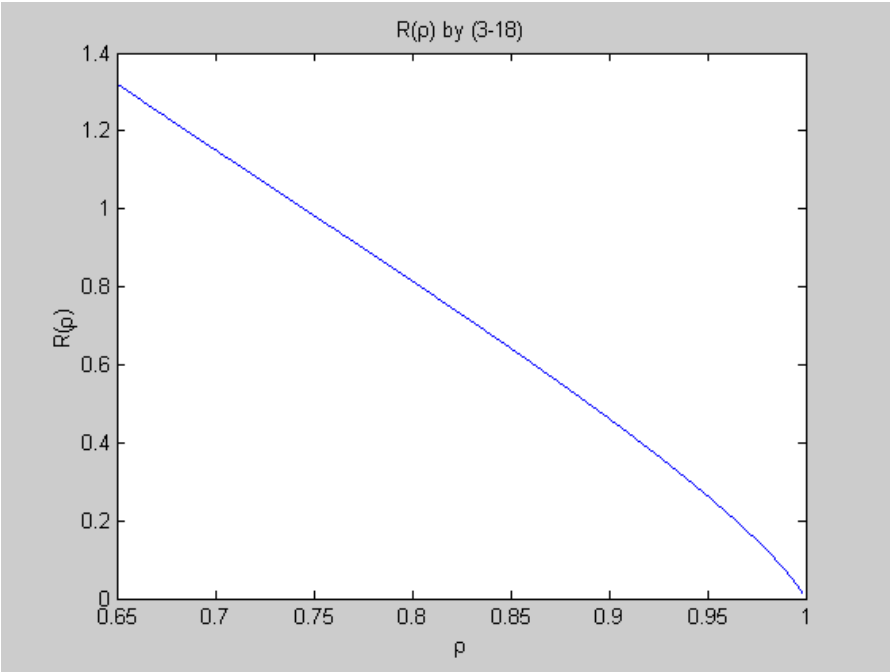

Figure 3-7 Relation between R and ρ by (3-18)

4.0 RATE CONTROL BASED ON LINEAR SOURCE MODEL

4.1 Introduction

The linear source model was developed in the last chapter. A rate control scheme is proposed in this chapter. The rate control scheme is designed to perform at three levels: Frame-Layer, Sub-frame Layer and Macro Block Layer.

The frame layer rate control decides how many bits are assigned to a certain frame based on the fullness of the buffer. The sub-frame layer rate control determines how many bits with minimum distortion are allocated to each of the decimated images given the number of bits of the whole image, depending on the complexity of the sub-images. The MB layer rate control is an adaptive rate control scheme to estimate some parameters and to achieve a certain target bit rate for a sub-frame by tuning the quantization step of each macro block.

A robust and accurate rate control is achieved by this layered control method. Details of the rate control algorithm of these three layers are discussed in this chapter. A summary of the overall rate control scheme is given in Section 4.5.

4.2 Frame Layer Rate Control

In this section, rate control of frame layer is discussed. Let C and F be the channel and frame rate, respectively. Let the frame skip threshold be M (Once the number of bits in buffer

reaches *M*, frame skipping is applied), which is set to be $\frac{C}{F}$ by default here. And let *W* denote the number of bits in the buffer.

Before the current frame is coded, we update the number of bits in the encoder buffer. According to [1],

$$W = \max(W_{prev} + B - \frac{C}{F}, 0) \qquad (4\text{-}1),$$

where *B* is the actual number of bits used for encoding the present frame, W_{prev} is the previous number of bits in the buffer. Initially, $W_{prev} = 0$, and *B* is the number of bits for the first intra frame. The updated number of bits in the buffer equals the previous buffer fullness W_{prev}, plus the bits used for the previous frame, and minus the number bits sent out by the channel per frame interval $\frac{C}{F}$. Of course the fullness of the buffer is larger than zero.

4.2.1 Frame-Skipping Scheme

If *W* is larger than or equal to *M*, the encoder skips encoding frames until the buffer fullness is below *M*. For each skipped frame, buffer fullness is reduced by an additional $\frac{C}{F}$ bits. Therefore, if the encoder needs to skip *L* frames, the updated buffer fullness is

$$W = \max(W_{prev} + B - (L+1)\frac{C}{F}, 0) \qquad (4\text{-}2).$$

Moreover, the buffer size should be larger than the frame-skipping threshold *M* to avoid buffer overflow.

The W bits in the encoder buffer belong to previously encoded frames and must be sent before those of the current frame, thus introducing a delay in the system denoted as buffer delay. With the frame skipping scheme, the number of bits in the buffer will not be larger than M at the beginning of the current frame's interval, and hence the buffer delay will be less than $\frac{M}{C}$ seconds. By default, we set $M=\frac{C}{F}$, and hence the maximum buffer delay is

$$\frac{M}{C}=\frac{1}{F} \qquad (4\text{-}3),$$

that is, the time of a frame interval [1]. In practice, the threshold M could be increased or decreased if we want a larger or smaller maximum buffer delay.

One thing that should be noted is that the frame-skipping scheme is only applied to inter frames. Intra frames are never skipped, since intra frames are used to eliminate the error caused by inaccurate motion prediction and thus to guarantee the quality of the sequence.

4.2.2 Target Bit Rate Decision

Now comes the problem of how to decide the target number of bits B_T for the current frame to be encoded. Here we use the method in TMN8 to decide the B_T [1].

$$B_T=\frac{C}{F}-\Delta \qquad (4\text{-}4),$$

where Δ is defined as follow:

$$\Delta=\begin{cases}\frac{W}{F}, for W > Z\cdot M\\ W-Z\cdot M, otherwise\end{cases} \qquad (4\text{-}5)$$

in which by default we set $Z=0.1$.

Thus Δ is a small value that provides feedback from the buffer fullness W. If W is larger than 10% of the maximum M, the frame target rate B_T is slightly decreased. Otherwise, B_T is slightly increased. Since $B \approx B_T$, the buffer fullness is according to (4-1) $W \approx \max(W_{prev} - \Delta, 0)$ and, as a result, the Δ correction helps maintain a small number of bits in the buffer without underflowing the buffer.

In most cases, the value of Δ is small, and hence the method selects a frame target B_T that is nearly constant throughout the video sequence. Since the complexity of the video sequence may change with time, this means that the quality of the encoded video will also vary. Nevertheless, having a near-constant target B_T is quite common for low-delay rate control, because, since M is small, larger fluctuations of B_T would produce large fluctuations of buffer fullness W that could easily result in undesirable frame skipping or buffer underflow.

The rate control scheme of frame layer is summarized in Figure 4-1.

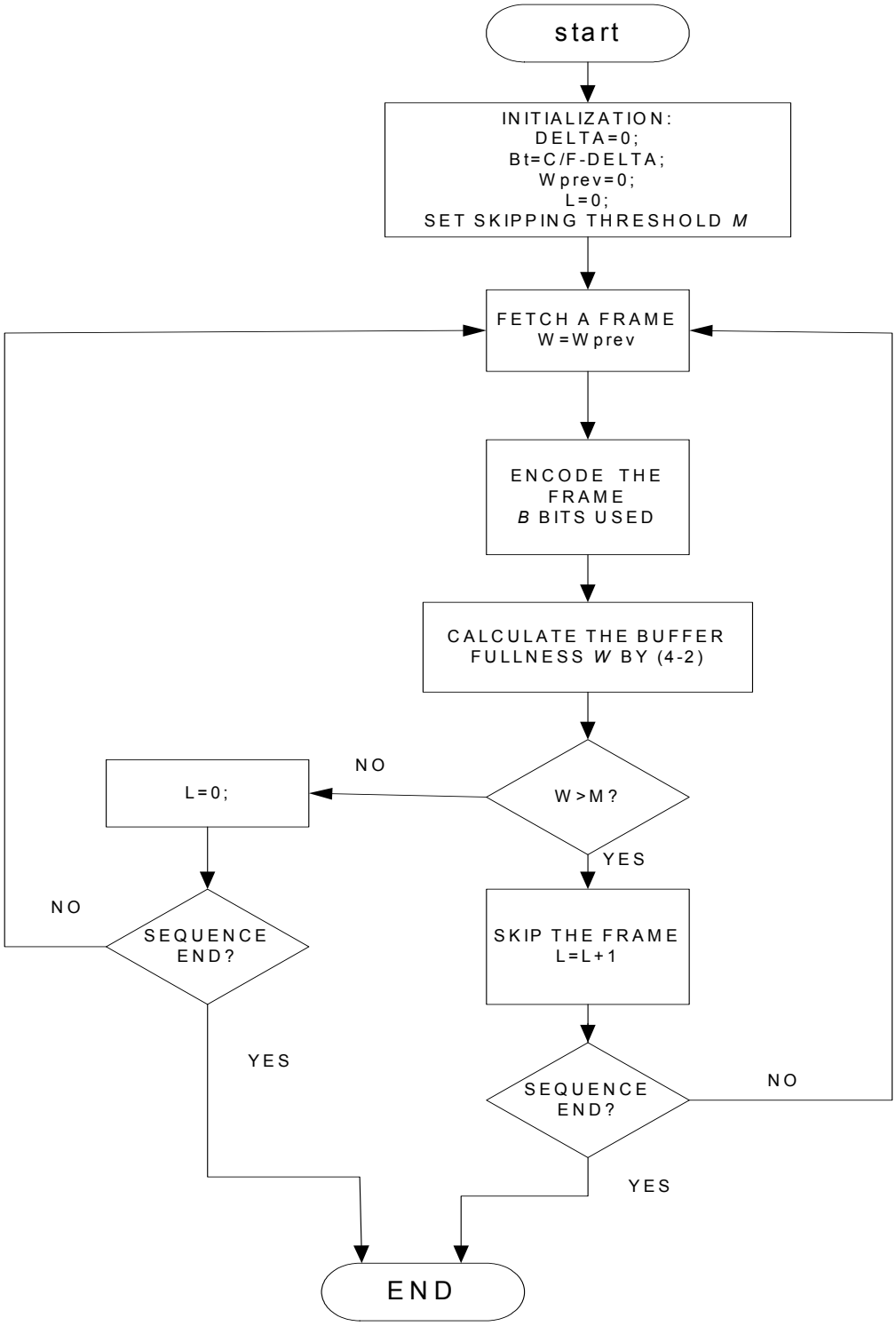

Figure 4-1 Frame Layer Rate Control

4.3 Sub-frame Layer Rate Control

Once the bit count assigned to a certain composited image is decided, rate control at sub-frame layer should be carried out to determine how to allocate a number of bits to each of the four sub-images. The problem turns out to be, given a certain number of bits and four sub-images, how to distribute the budgeted bits among four sub-images to achieve best overall quality. That is, if D_i , $i = 1,2,3,4$, is the distortion of *ith* sub-image, defined by the absolute value error, R_i is the bit rate of the *ith* sub-image and $R_T = {B_T}/{numberofcoefficients}$ is the average bit count per coefficient, then $D_i = D_i(R_i)$, and $\sum_{j=1}^{4} R_j = R_T$, $R_i > 0$, and we try to find out R_i to minimize $\sum_{j=1}^{4} D_j$, where $i = 1,2,3,4$.

In [12], an optimal bit allocation method for video coding is given. The author develops an approximate distortion model and uses Lagrange method to do the optimization. There are two problems involved in this method. First, the model is not accurate enough to describe the distortion behavior. Second, Lagrange method does not guarantee that the optimized $R_i > 0$.

In this section, we develop a distortion model and an optimum bit allocation scheme similar to the one in [12] on the level of sub-frame.

4.3.1 Estimation of Distortion Model $D(R)$

Since the loss caused by the decimation is very low in the DCT domain, we use the original image to estimate the distortion model of the decimated image. We assume that the

distortion model of the original image and that of the decimated image are approximately the same.

The distortion of an intra frame and inter frame in the ρ- domain is given by (3-12) and (3-17) respectively.

$$D(\rho) = \begin{cases} \frac{1}{\lambda} \cdot \frac{1-t}{1+t} \\ \frac{1}{\lambda}\left(\frac{1-2t+t^{2a}-2t(a-1)\log(t)}{1-t^{2a}} \right) \end{cases} \qquad (4\text{-}6).$$

where the first equation applies to intra-frame, and the second to inter-frame. From Chapter 3, we already know the linear relation between ρ and bit rate *R*, that is,

$$R = \theta \cdot t \qquad (4\text{-}7)$$

where in equations (4-6) and (4-7) $t = 1-\rho,$ and $a = \frac{0.5Q}{0.5Q+\alpha} = \frac{2}{3}$, ($\alpha = 0.25Q$), and Q is the quantization step.

Equation (4-6) is a complex expression that is hard to analyze, we use a simple exponential function to approximate it:

$$D(\rho) = \frac{1}{\lambda} e^{-\beta \cdot (1-\rho)^{\gamma}} \qquad (4\text{-}8),$$

where β and γ are parameters to be decided. γ is determined by trial-and-error. By setting for intra frame $\gamma = 1$, and for inter frame, $\gamma = 0.85$, we can then choose a value for β that provides an accurate approximation for both intra frame case and inter frame. Figure 4-2 and Figure 4-3 shows how (4-8) matches (4-6) when $t = 1-\rho$ is small, i.e. low bit rate video.

Combining (4-7) and (4-8), we can get the distortion model $D(R)$:

$$D(R) = \frac{1}{\lambda} e^{-(\beta/\theta^{\gamma}) \cdot R^{\gamma}} \qquad (4\text{-}9)$$

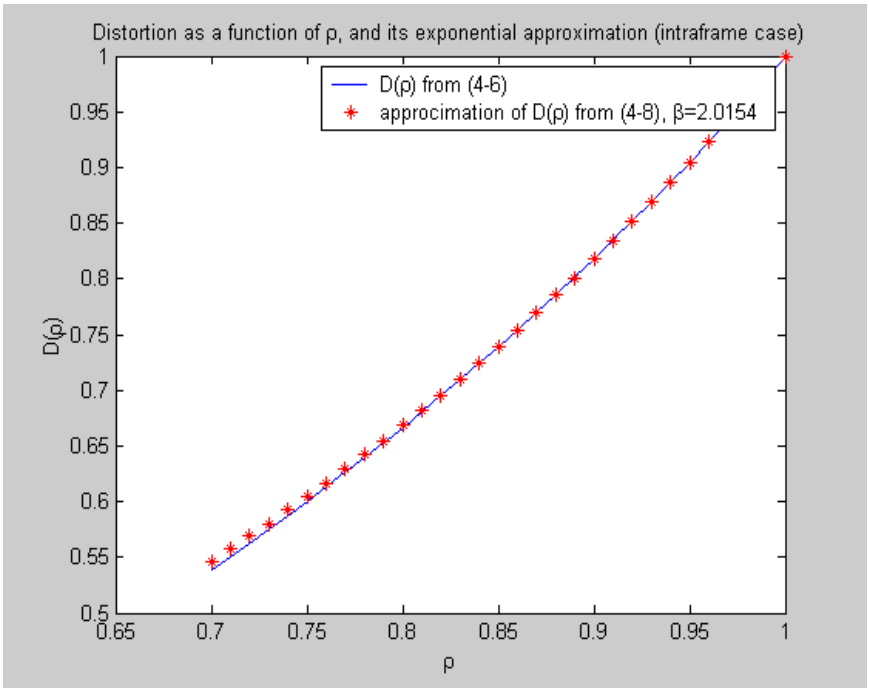

Figure 4-2Exponential Approximation of Distortion vs. ρ function (intraframe case)

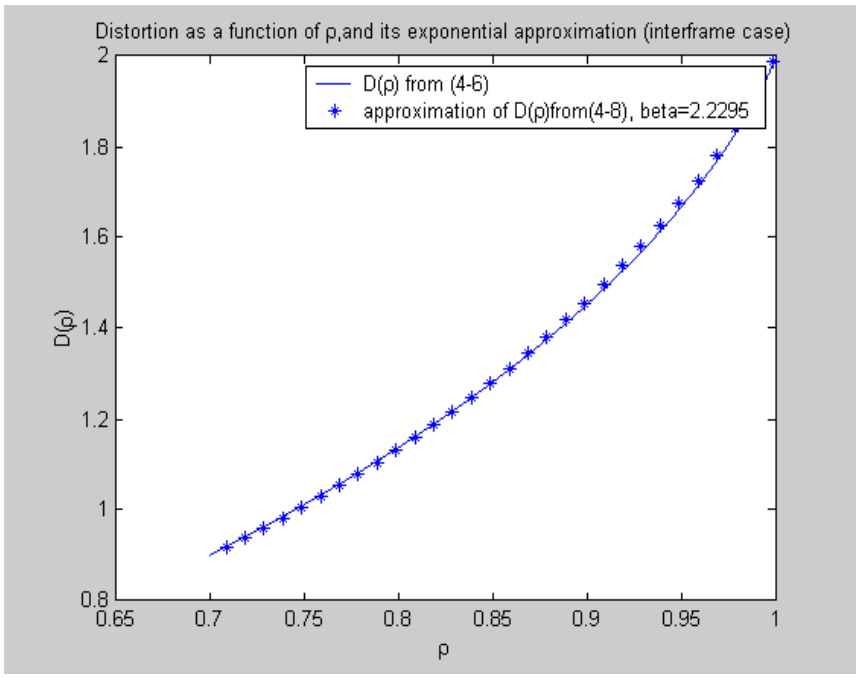

Figure 4-3 Exponential Approximation of Distortion vs. ρ function (interframe case)

4.3.2 Bit Allocation Optimization

For the distortion model developed in the last section, there are several parameters to be estimated: $\lambda_i, \beta_i, \theta_i, i = 1,2,3,4$ in equation (4-9). They can be estimated when transcoding the incoming frames.

When receiving the *ith* incoming frames ($i = 1,2,3,4$), the bits rate R_i (average number of bits for each pixel), and the percentage of zeros ρ_i are calculated; t_i can be calculated by $t_i = 1 - \rho_i$. θ_i can be estimated as

$$\hat{\theta}_i = \frac{R_i}{t_i} \qquad (4\text{-}10).$$

According to (3-10) and (3-16), λ is estimated by:

$$\hat{\lambda}_i = \begin{cases} \dfrac{-2\log(t_i)}{Q_i} \\ \dfrac{-\log(t_i)}{0.5Q_i + \alpha} \end{cases} \qquad (4\text{-}11).$$

where the first equation is for intra-frame and the second for inter-frame. D_i can be estimated by (4-6), and denoted by $\hat{D}_i$. Such that β_i can be calculated according to (4-9):

$$\hat{\beta}_i = \left(\frac{\hat{\theta}_i}{R_i}\right)^{\gamma} \log\left(\frac{1}{\hat{\lambda}_i \hat{D}_i}\right) \qquad (4\text{-}12).$$

Now that all the parameters are known and the distortion model is developed, the optimization problem turns out to be a constrained nonlinear optimization problem: That is, given $\hat{D}_i(R_i) = \frac{1}{\hat{\lambda}_i} e^{-(\hat{\beta}_i / \hat{\theta}_i^{\gamma}) \cdot R_i^{\gamma}}$ and $\sum_{j=1}^{4} R_j = R_T$, $R_i > 0, i = 1,2,3,4$, find out $R_{i,}$ to minimize

$\sum_{j=1}^{4} \hat{D}_j$. There are a lot of algorithms to solve this kind of problem with low computational complexity, such as Sequential Quadratic Programming (SQP) method, which are beyond the scope of this thesis. We used the function *fmincon* in Matlab to solve it.

4.4 MB Layer Rate Control

When the target bit budget is decided for each sub-image, a rate control method on Macro Block layer is performed to precisely code the sub-frame using the target bit count by adjusting the quantization step of each macro block.

4.4.1 Adaptive Estimation of θ

Let N_m be the number of the MBs where have already been encoded in the current sub-frame. In a 16×16 MB of a black-and-white image, there are 256 coefficients in total. Let R_m be the number of bits used to encode these N_m MBs. We denote the number of zeros in these MBs by ρ_m. Based on the linear model, multiplying both sides of equation (3-7) with $256\times N_m$ ($16\times16\times N_m$, the number of pixels of N_m macroblocks), The parameter θ can be estimated as follows:

$$\bar{\theta} = \frac{R_m}{256\times N_m - \rho_m} \qquad \text{(4-13).}$$

The estimated $\bar{\theta}$ is then applied to the current macroblock, so that the estimated $\bar{\theta}$ is an accumulative statistic of the coded MBs. As the coding goes on, the estimated $\bar{\theta}$ should

converge to the true value of θ of the current sub-frame. Figure 4-4 shows the convergence of $\overline{\theta}$.

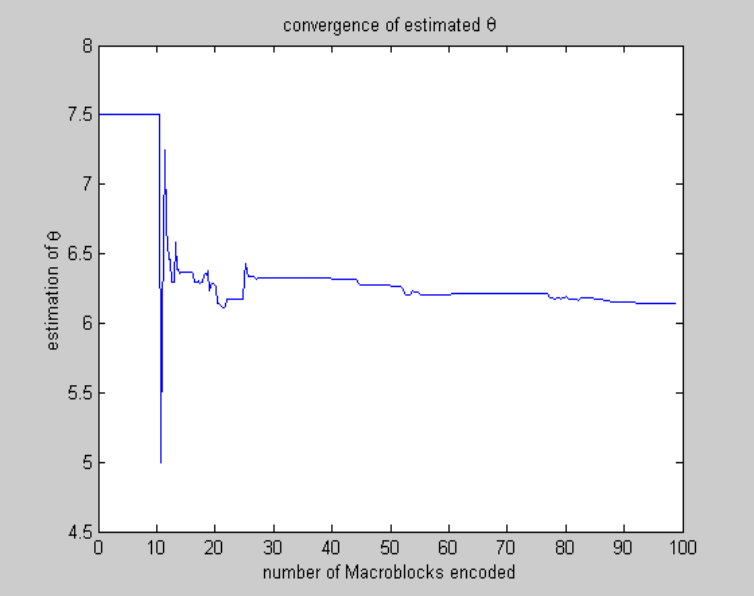

Figure 4-4 Convergence of $\overline{\theta}$ (Encoded image is "Claire")

4.4.2 Rate Control Algorithm

Based on the adaptive estimation of θ, the proposed rate control algorithm is shown below. Note that the rate control algorithm is only to control the bit rate of DCT coefficients. Bits of MVs and overhead are not in consideration here.

Step 1 *Initialization*:

Before encoding the first MB, set $N_m = R_m = \rho_m = 0$. Compute the distribution of the DCT coefficients to develop the $Q - \rho$ table for further use. Set $\bar{\theta} = 7.5$, which is the average value for typical video sequences.

Step 2 *Find Q*:

According to (4-13) and the linear source model (3-7), the number of zeros to be produced by quantizing the rest MBs should be

$$\hat{\rho} = 256 \times (N - N_m) - \frac{R_i - R_m}{\bar{\theta}} \qquad (4\text{-}14),$$

where *N* is the total number of the macro blocks in the *ith* sub-frame, R_i is the target number of bits allocated to the *ith* sub-frame. Based on the one-to-one mapping between Q and ρ, the quantization parameter is determined. The current MB is then quantized with *Q* and encoded.

Step 3 *Update*:

Let ρ_0 and R_0 be the number of zeros and number of bits produced by the current MB, respectively. Set $\rho_m = \rho_m + \rho_0$, $R_m = R_m + R_0$ and $N_m = N_m + 1$. If $N_m \geq 10$, update the value of $\bar{\theta}$ according to (4-13).

Step4 *Loop*:

Repeat step 2 and step 3 for the next MB until all the MBs in the current sub-frame are encoded.

4.4.3 Modifications

It is observed that the errors given by this rate control algorithm performed on the sub-images in the same sequence is highly related. That is, for the sub-images in the same sequence

to be encoded, the bit count resulted from the rate control method is probably always above or below the target bit count by a nearly constant number. This is because of the similarity of the sub-images in the same sequence.

To get rid of this nearly constant error, a parameter *e* is set to correct the error. Initially set $e_0 = 0$, and update *e* at the end of encoding each sub-frame:

$$e_i = w \cdot error_i + (1 - w) \cdot e_{i-1} \qquad (4\text{-}15),$$

where $error_i$ is the error caused by encoding the *i*th sub-frame (the difference between target bit number and the bit number achieved by the encoding), and *w* is a weight. Set *w* be 0.5 in our experiment.

4.5 Summary Of Rate Control Method

The overall rate control scheme is shown is Figure 4-5. This rate control algorithm has low computational complexity and implementation cost and it performs better than TMN8 for our Transcoding-Compositing system. The experiment results are given in the next chapter.

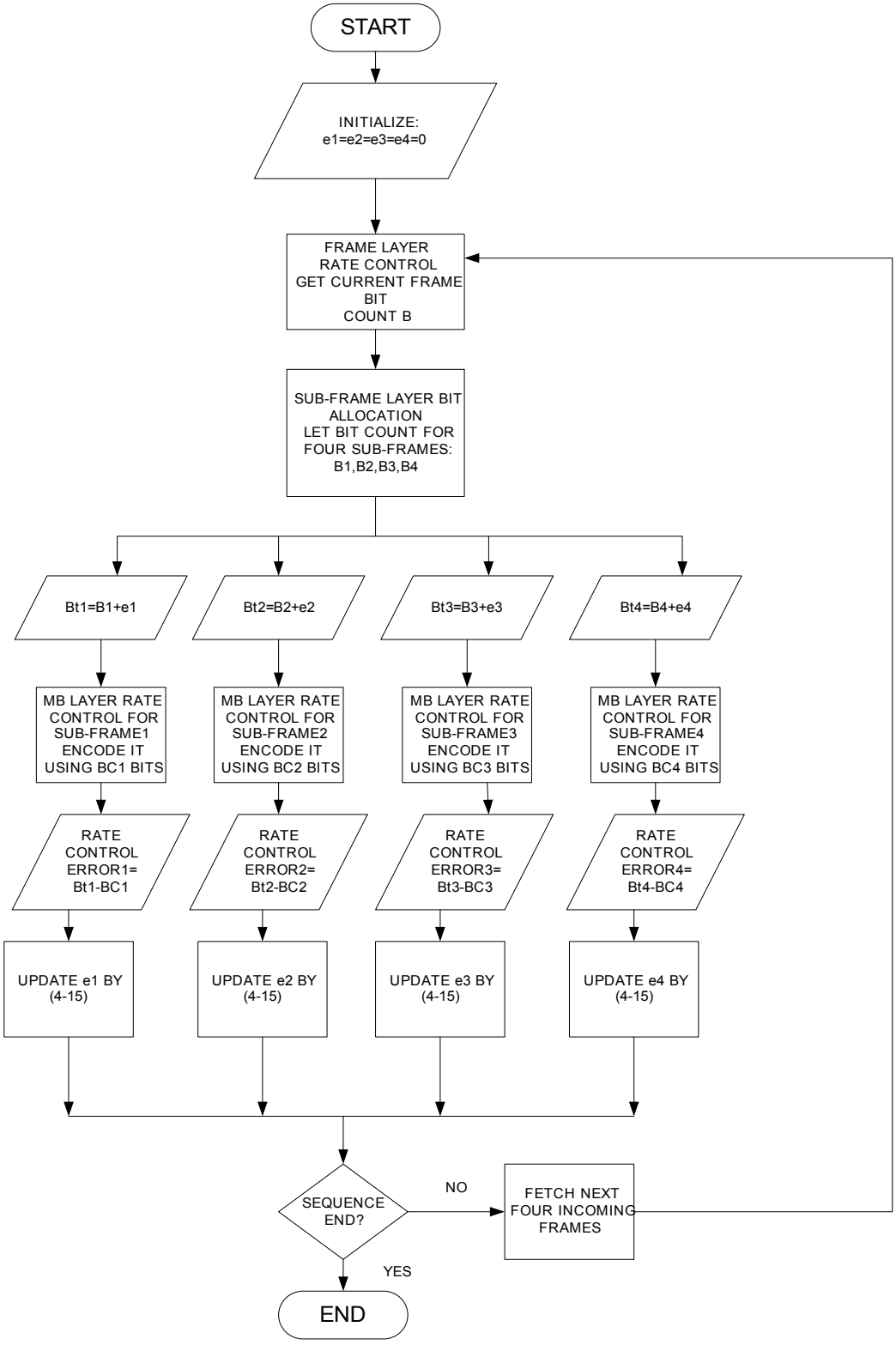

Figure 4-5 Overall Rate Control Scheme

5.0 EXPEREMENTAL RESULTS

It has been shown in [11] that the linear source model in ρ domain has a much better performance over the quadratic model employed by TM5. To make the comparison more meaningful, We have compared the proposed three-level rate control scheme with the rate control scheme in TMN8, which is regarded as the best control in use. Experiments have shown that our rate control scheme is an improvement over TMN8 in terms of control accuracy and video PSNR for composited images.

First of all, we tested the performance of the Macroblock Layer rate control by applying MB-Layer rate control on a single image instead of a composited image. we chose the sequence "Claire", a typical "head-shoulder" video, as the sample sequence. The video sequence form is CIF, i.e. each frame has 288x352 pixels. Since it is most meaningful to control interframes' bit rates in reality, only interframes are tested. The channel rate C is set to be 48000bit/s, and frame rate F=10frame/s, so the target bit count for each frame is around 4800 bits plus a small adjustable quantity based on the buffer fullness. Assume that the buffer size is 4800bits. The buffer fullness is calculated by (4-2). Frame skipping is not considered in this experiment. Figure 5-1 shows the control accuracy for both proposed method and the TMN8 method. The continuous line stands for the target bit count and the stars are controlled bit count of each method. It can be seen that the controlled bit count of proposed method is closer to the target bit count (denoted by solid line), which means more accurate control, while in the case of TMN8, several stars are much above the target line. The algorithms used in TMN8 are from [1].

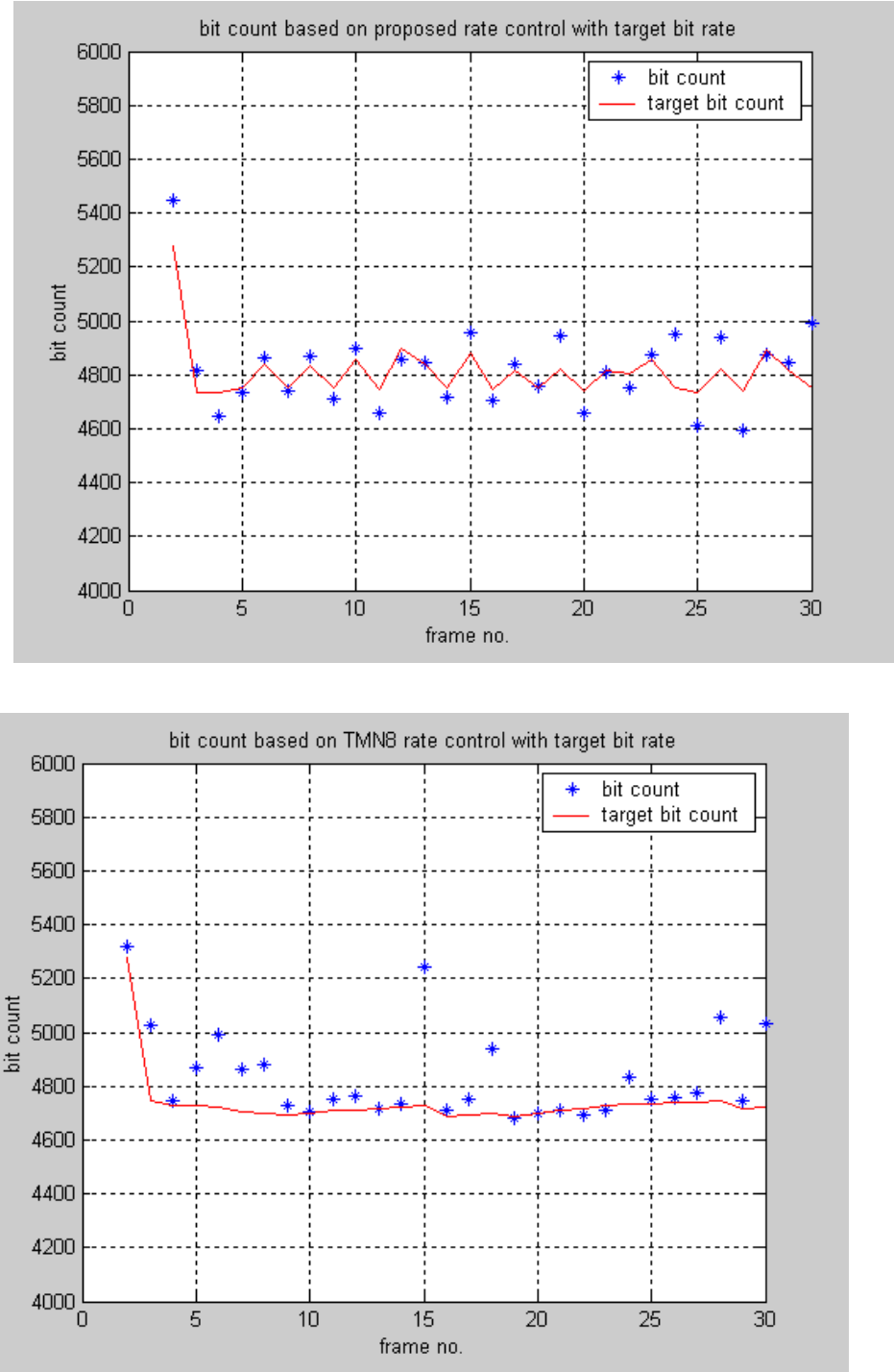

Figure 5-1 Proposed MB Layer Rate and TMN8 Rate Control

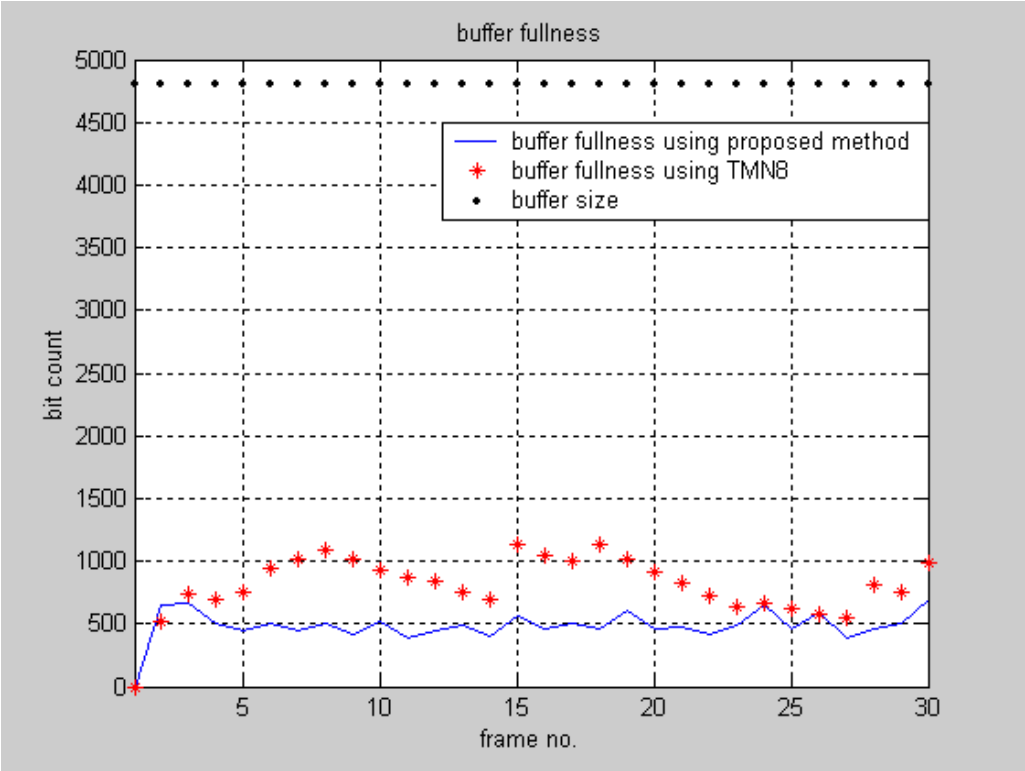

Figure 5-2 Buffer Fullness Comparison between Proposed Control and TMN8 Control

Table 5-1 Control Error Comparison

Gross Target Bit Count	Control Error Based on Proposed MB Layer Rate Control	Control Error Based on TMN8 Rate Control
4800bit/frame	1.37%	2.03%
4200bit/frame	1.63%	1.75%
3600bit/frame	1.14%	1.40%
3000bit/frame	3.02%	3.04%
2400bit/frame	4.77%	5.03%

Figure 5-2 shows the buffer fullness of the two methods. The solid line is the buffer fullness of the proposed method and the stars represent buffer fullness of TMN8 control. For most of the time, the buffer fullness of the proposed method is below that of TMN8 control. There are consistently about 500 bits in the buffer with the proposed control, always below 20% of the buffer size. If frame skipping is applied, control with the proposed method will be subjected to less skipping, avoiding inconsistent video quality and unnatural movement.

Table 5-1 shows the control error comparison of the proposed method and the TMN8 method. Control error is calculated by

$$error = \frac{|B_T - B|}{B_T} \times 100\% \qquad (5\text{-}1),$$

where B_T is the target bit count and B is the resulting bit count. Comparisons are made on several gross target bit count. Actual target bit count is equal to gross target bit count with a small variation. Generally speaking, the control error increases when the target bit count is decreased in both controls. This reflects the fact that it is more difficult to make accurate control when the target bit count is getting smaller. Even with target bit count as low as 2400bit/frame, the control error is less than 5% with our proposed method. And when the target bit is about 4800bit/frame, the control accuracy is above 98%. It can be seen from the table that the proposed method outperforms the TMN8 MB-layer control by 0.3% on average.

Table 5-2 Sub-frame Bit Allocation

Frame No.		Claire	Miss America	Salesman	Trev	Total bit count
1 (intra frame)	σ	150.2465	88.5175	102.6912	107.0698	50000bits
	Target Bit Count	16749	9867	11448	11936	
	PSNR (dB)	37.0816	39.2686	35.5886	37.5730	
2 (inter)	σ	4.6787	4.8387	7.0235	6.4273	5000bits
	Target Bit Count	601	604	1956	1839	
	PSNR (dB)	38.5838	39.9804	40.7137	39.3874	
3 (inter)	σ	4.4339	4.2239	5.9752	5.4027	5000bits
	Target Bit Count	706	654	1891	1749	
	PSNR (dB)	36.1530	41.2575	41.6835	39.0411	
4 (inter)	σ	4.1158	4.1825	6.0494	5.1720	5000bits
	Target Bit Count	654	671	1949	1726	
	PSNR (dB)	40.9031	40.6767	40.9540	38.9186	
5 (inter)	σ	4.1693	4.3301	6.2481	5.6191	5000bits
	Target Bit Count	623	663	1934	1780	
	PSNR (dB)	42.3739	41.0910	39.2352	37.9609	
6 (inter)	σ	4.0833	4.5365	6.3267	5.7153	5000bits
	Target Bit Count	588	697	1931	1784	
	PSNR (dB)	39.8790	39.4765	38.5982	38.3740	
7 (inter)	σ	3.9853	4.7864	6.6220	5.4800	5000bits
	Target Bit Count	554	746	1986	1714	
	PSNR (dB)	42.5237	40.9339	39.6333	39.4890	
8 (inter)	σ	4.0548	4.2428	7.0352	5.5333	5000bits
	Target Bit Count	571	617	2086	1726	
	PSNR (dB)	41.9259	42.8762	37.6546	40.1403	
9 (inter)	σ	4.1621	4.4086	7.2523	5.2246	5000bits
	Target Bit Count	588	647	2123	1642	
	PSNR (dB)	39.4460	39.2661	35.7529	41.2074	
10 (inter)	σ	3.7461	4.4044	7.3032	5.4611	5000bits
	Target Bit Count	495	652	2145	1708	
	PSNR (dB)	41.9159	42.1380	35.5974	40.1429	
11 (inter)	σ	3.6800	4.3264	7.3049	5.4515	5000bits
	Target Bit Count	488	641	2159	1712	
	PSNR (dB)	42.4417	41.7181	34.6674	39.4175	
12 (inter)	σ	3.9104	4.3549	7.1556	5.7720	5000bits
	Target Bit Count	522	627	2088	1783	
	PSNR (dB)	41.6416	40.8517	34.6811	37.7981	
13 (inter)	σ	3.4247	4.3194	7.2249	6.3064	5000bits
	Target Bit Count	428	615	2097	1860	
	PSNR (dB)	40.8723	40.9517	33.6542	36.1123	

Now let us consider composited frames. The CIF frames are tested, each one contains four decimated sub-images of 144x176 pixels. We chose “Claire”, “Miss America”, “Trev”, and “Salesman” as four incoming streams, all of which are CIF images and subjected to decimation by 2. Instead of assigning the bit count evenly; sub-frame layer bit allocation decides how many bits should be assigned to each sub-image to achieve the least distortion. Table 5-2 shows the results of the bit allocation on 13 frames. The first frame is intraframe and the rest are interframes. Suppose the intraframe is to be coded with 50,000 bits and the interframes with 5000bits. The estimation of standard deviation σ of each sub-image, the resulted target bit count of each sub-frame, and their PSNRs after encoding-decoding are listed.

As has been shown in Chapter 4, bit allocation is based on the source rate-distortion model with several parameters such as standard deviation σ, and the slope θ of the linear $R-\rho$ model, etc. All these parameters represent the source image feature. To make the table concise, only the standard deviation σ is shown for reference. Large σ means high image complexity in most cases. More bits are needed to encoded images with high complexity. Experiments prove this rule. More bits are allocated to “Trev” and “Salesman”, which are much more complex videos than “Claire” and “Miss America”. According to our proposed bit allocation method, the number of bits assigned to “Salesman” is 3 to 4 times of that assigned to “Claire”.

The PSNRs of the encoded sub-images are also given to demonstrate the effect of bit allocation. All the encoded sub-images have PSNRs higher than 30dB. For the simple images “Claire” and “Miss America”, the PSNRs are above 36dB, and even above 40dB in some cases. The PSNRs of “Trev” and “Salesman” decline obviously because of the accumulated quantization error. However, they are still around 35dB. It can be concluded that images like “Salesman” are relatively hard to compress compared to other sub-images.

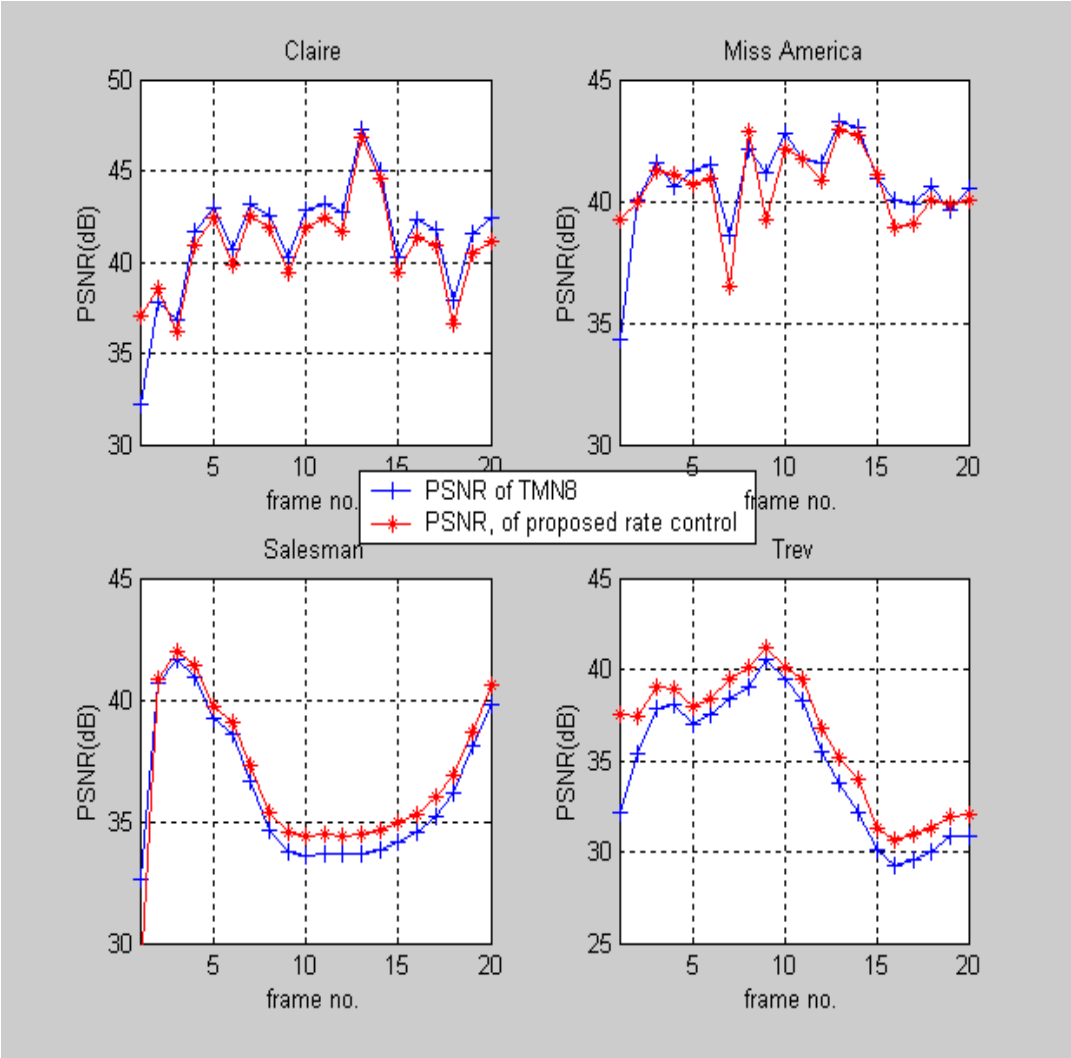

Figure 5-3 PSNR Comparison between TMN8 and Proposed Rate Control

Figure 5-3 shows the PSNR Comparison between TMN8 rate control and the proposed rate control. Again we chose the same four incoming sequences, all of which have one intraframe each followed by 19 interframes. We processed the incoming sequences with the DCT-based Transcoding-Compositing system with the proposed rate control scheme and got PSNR values for the four sub-images respectively. Also we processed the incoming sequences with TMN8 rate control. No bit allocation is used and all the sub-images share evenly the frame bit budget. Again we got the PSNR for them. In Figure 5-3, the red line represents the PSNR value of the proposed rate control while the blue line stands for the PSNR value of TMN rate control.

Among the four incoming sequences, “Claire” and “Miss America” are the two relatively easy to be compressed. Their PSNRs are around 40dB as shown in the figure. However, “Trev” and “Salesman” are more complex with a PSNR of 35dB on average. As shown in Figure 5-3, our proposed rate control raise the PSNRs of “Trev” and “Salesman” by 0.5~1dB by compromising in encoding “Claire” and “Miss America”. Trade-off always exists in assigning bit budget to four sub-images; our goal is to raise the overall image quality by adjusting the qualities among these sub-images. Figure 5-3 demonstrates the bit allocation effect. By using the proposed rate control, PSNR of all sub-frames are kept above 30dB.

6.0 CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

Rate control in video communication has been explored extensively in recently years. To develop a proper rate control scheme for real-time multimedia video conferencing, we considered the DCT-based transcoding-compositing system. We made some assumptions to simplify the problem. In order to achieve low bit rate communication, we proposed a KL-based scene change detection algorithm to decide frame mode for composited image in Chapter 2. Then a three-layer rate control scheme is developed to obtain both desirable video quality and accurate rate control in Chapter 4.

The proposed rate control method is based on a newly established linear source model, which is discussed in detail in Chapter 3. Instead of modeling in Q domain, the percentage of zeros in the quantized image, which can be easily calculated, is chosen to be the modeling variable. This simple and accurate model paved the way for the new rate control scheme.

Chapter 5 shows the experimental results of the proposed rate control scheme. Experiments haven shown that the proposed rate control has at least comparable performance of TMN8, and outperforms over TMN8 in some cases. Sub-frame bit allocation guarantees that the bits are smartly assigned and overall distortions are kept at a low level. Experimental results prove that the proposed rate control can meet the goal of encoding the video in target low bit rate precisely as well as achieving good video quality.

6.2 Future Research

One of the future research steps is to explore the theoretic basis beyond the linear source model. The justification for the linear model given in this thesis is not given in a formal style. Further research in the linear model will help us to have a better understanding of images or video signals. Another direction of future research is to apply the linear model to other coded images, such as MPEG-2, Wavelet Transform based coded images. And we also need to make further modifications to perfect the proposed rate control method.

APPENDICES

APPENDIX A

In H.263, there are six types of frame: Intra or I Frame, Inter or P Frame, PB Frame and Improved PB Frame, B Frame, EI Frame, EP Frame. Only first two types are concerned in this thesis. The coding mode using prediction is called inter, the coding mode using no prediction is called intra. Intra coding is signaled at the picture level (I frame for intra or P frame for inter) by Picture Coding Type in the field of PTYPE (Type Information), and at the macroblock level in P frames by Macroblock Type. The macroblock types are listed in the Table 6-1.

Table 6-1 H.263 Macroblock Types and Included Date for Normal Frames

Frame Type	MB Type	Name
Inter	Not coded	-
Inter	0	Inter
Inter	1	Inter + q
Inter	2	Inter 4v
Inter	3	Intra
Inter	4	Intra + q
Inter	5	Inter 4v + q
Inter	Stuffing	-
Intra	3	Intra
Intra	4	Intra + q
Intra	Stuffing	-

APPENDIX B

A detailed derivation of (3-11) and (3-17) is given in this appendix. Note that

$$\int_s^t \lambda e^{-\lambda x} dx = e^{-\lambda s} - e^{-\lambda t}$$

and

$$\int_s^t \lambda x e^{-\lambda x} dx = \left[s e^{-\lambda s} - t e^{-\lambda t} \right] + \frac{1}{\lambda} \left[e^{-\lambda s} - e^{-\lambda t} \right].$$

From (3-9),

$$D(Q) \approx \int_{-0.5Q}^{0.5Q} p_l(x) x dx + \sum_{\substack{i=-\infty \\ i \neq 0}}^{\infty} \int_{(i-0.5)Q}^{(i+0.5)Q} p_l(x) \mid x - iQ \mid dx$$

$$= 2 \int_0^{0.5Q} p_l(x) x dx + 2 \sum_{i=1}^{\infty} \int_{(i-0.5)Q}^{(i+0.5)Q} p_l(x) \mid x - iQ \mid dx$$

$$= \int_0^{0.5Q} \lambda e^{-\lambda x} x dx + \sum_{i=1}^{\infty} \int_{(i-0.5)Q}^{(i+0.5)Q} \lambda e^{-\lambda x} \mid x - iQ \mid dx \qquad \text{Let } x - iQ = u$$

$$= \left[-0.5Q\lambda + \frac{1}{\lambda}(1 - e^{-0.5Q\lambda}) \right] + \sum_{i=1}^{\infty} \int_{-0.5Q}^{0.5Q} \lambda e^{-\lambda(u+iQ)} \mid u \mid du$$

$$= \left[-0.5Q\lambda + \frac{1}{\lambda}(1 - e^{-0.5Q\lambda}) \right] + \sum_{i=1}^{\infty} e^{-\lambda i Q} \left[\int_{-0.5Q}^{0.5Q} \lambda e^{-\lambda u} \mid u \mid du \right]$$

$$= \frac{1}{\lambda} \left[1 + \frac{e^{-\lambda Q}}{1 - e^{-\lambda Q}} (2 - e^{-0.5\lambda Q} - e^{0.5\lambda Q}) - e^{-0.5\lambda Q} \right]$$

From (3-15),

$$D(Q) \approx 2\int_0^{0.5Q+\alpha} p_l(x)x\,dx + 2\sum_{i=1}^{\infty}\int_{(i-0.5)Q+\alpha}^{(i+0.5)Q+\alpha} p_l(x)\,|\,x-iQ\,|\,dx \quad \text{Let } x-iQ=u$$

$$= \int_0^{0.5Q+\alpha} \lambda e^{-\lambda x}x\,dx + \sum_{i=1}^{\infty}\int_{-0.5Q+\alpha}^{0.5Q+\alpha} \lambda e^{-\lambda(u+iQ)}\,|\,u\,|\,du$$

$$= -(0.5Q+\alpha)e^{-(0.5Q+\alpha)\lambda} + \frac{1}{\lambda}\left(1-e^{-(0.5Q+\alpha)\lambda}\right) + \sum_{i=1}^{\infty} e^{-i\lambda Q}\left[-\int_{-0.5Q+\alpha}^{0}\lambda e^{-\lambda u}u\,du + \int_0^{0.5Q+\alpha}\lambda e^{-\lambda u}u\,du\right]$$

$$= \frac{1}{\lambda}\left[1-e^{-(0.5Q+\alpha)\lambda} + \frac{e^{-Q\lambda}}{1-e^{-Q\lambda}}\left(2-e^{-(0.5Q+\alpha)\lambda}-e^{(0.5Q-\alpha)\lambda}\right) - e^{-(0.5Q+\alpha)\lambda}\frac{2\alpha\lambda}{1-e^{-Q\lambda}}\right]$$

Let $a = \dfrac{0.5Q}{0.5Q+\alpha}$, $\quad t = 1-\rho$, $\quad$ and $\quad \alpha = 0.25Q$,

$t^a = e^{0.5Q\lambda}$, and $\alpha\lambda = (a-1)\log(t)$,

$$\therefore D(t) = \frac{1}{\lambda}\left(\frac{1-2t+t^{2a}-2t(a-1)\log(t)}{1-t^{2a}}\right).$$

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Jordi Ribas-Corbera, "Rate Control in DCT Video Coding for Low-Delay Communications", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No.1, pp.172-185, Feb. 1999

[2] Wei Ding, "Rate Control of MPEG Video Coding and Recording by Rate-Quantization Modeling", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 6, No.1, pp.12-20, Feb. 1996

[3] L.F.Chaparro, C.C.Li, "New Multi-point Video Conferencing", Final Technical Report to the Pittsburgh Digital Greenhouse, April 2003

[4] T.H. Kweh, "Rate control algorithm for block-based variable rate video encoders", Electronics Letters, 4th, Vol.32, No.14, pp.1277-1278, July, 1996

[5] Boon-Lock Yeo, "Rapid Scene Analysis on Compressed Video", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 5, No.6, pp.533-541, Dec. 1995

[6] Chung-Lin Huang, "A Robust Scene-Change Detection Method for Video Segmentation", IEEE Transactions on Circuits and Systems for Video Technology, Vol.11, No.12, pp.1281-1288, Dec. 2001

[7] Oliver Bao, "Scene Change Detection Using DC Coefficients", pp.II-421-II-424, IEEE International Conference on Image Processing 2001

[8] R. A. Devijver and J. Kittler, *Pattern Recognition: A statistical approach*. London: Prentice Hall, 1982

[9] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, NJ: Prentice Hall, 1984

[10] Tihao Chiang, "A New Rate Control Scheme Using Quadratic Rate Distortion Model", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 7, No.1, pp.246-250, Feb. 1997

[11] Zhihai He, "A Linear Source Model and a Unified Rate Control Algorithm for DCT Video Coding", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No.11, pp.970-982, Nov. 2002

[12] Zhihai He, “Optimum Bit Allocation and Accurate Rate Control for Video Coding via rou- Domain Source Modeling”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12, No.10, pp.840-849, Oct. 2002

[13] Zhihai He, “Optimal Bit Allocation for Low Bit Rate Video Streaming Applications”, pp.I-73-I-76, IEEE International Conference on Image Processing 2002

[14] Zhihai He, “Low-Delay Rate Control for DCT Video Coding via rou-Domain Source Modeling”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No.8, pp.928-940, Aug. 2001

[15] Zhihai He, “Novel Rate-distortion Analysis Framework for Bit Rate and Picture Quality Control in DCT Visual Coding”, IEE Processings.-Vis. Image Signal Process, Vol.148, No.6, pp.398-406, Dec. 2001

[16] Keith Jack, *Video Demystified, Ahandbook for the Digital Engineer, third edition*, LLH Technology Publishing, Eagle Rock, Virginia, 2001

[17] Jordi Ribas-Corbera, “A Frame-Layer Bit Allocation for H.263+”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No.7, pp.1154-1158, Oct. 2000

[18] Hung-Ju Lee, “Scalable Rate Control for MPEG-4 Video”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No.6, pp.878-894, Sep. 2000

[19] Eric C. Reed, “Constrained Bit-Rate Control for Very Low Bit-Rate Streaming-Video Applications”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No.7, pp.882-889, July. 2001

[20] I.-Ming Pao, “Encoding Stored Video for Streaming Applications”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No.2, pp.199-209, Feb. 2001

[21] Viresh Ratnakar, “An Efficient Algorithm for Optimizing DCT Quantization”, IEEE Transactions on Image Processing, Vol. 9, No.2, pp.267-270, Feb. 2000

[22] Chia-Wen Lin, “Dynamic Rate Control in Multipoint Video Transcoding”, IEEE International Symposium on Circuits and Systems, pp.II-17-II-20, May 28-31,2000, Geneva, Switzerland

[23] Deepak S. Turaga, “No Reference PSNR Estimation for Compressed Pictures”, pp.III-61-III-64, IEEE International Conference on Image Processing 2002

[24] Anthony Vetro, “Rate-Reduction Transcoding Design for Wireless Video Streaming”, pp.I-29-I-32, IEEE International Conference on Image Processing 2002