Scalable Video and Audio Techniques for Video Conferencing

by

Fung Kai Tat, BEng(Hons)

A thesis submitted for the Degree of Master of Philosophy in the Department of Electronic and Information Engineering of the Hong Kong Polytechnic University

Department of Electronic and Information Engineering
The Hong Kong Polytechnic University

September 2001

Abstract

With the advance of video and audio compression and networking technologies, networked multimedia services, such as multipoint video conferencing, video on demand and digital TV, are emerging. We envision a central server(MCU) that may have to support quality of service to heterogeneous clients or transmission channels and it is in this scenario that this server has the capability to perform transcoding in video and audio mixing.

In video transcoding, the conventional approach needs to decode the incoming video bitstream in the pixel domain, and the decoded video frame is re-encoded at the desired output bitrate according to the capability of the clients' devices and the available bandwidth of the network. This involves high processing complexity, memory, delay and video degradation. In the audio mixing, the audio signal is usually distorted by the background noise from other channels and makes the speech signal quality degraded.

The aim of this study is to find ways that can reduce the computational complexity and provide good quality of video and audio in the video conferencing. In this thesis, we focus on four major aspects of a video conferencing system. They are the video transcoding in multipoint video conferencing, the wavelet based video coder, speech recovery and audio coding. The first half of the thesis is concerned with the video processing while the second half is concerned with the audio processing.

In the first half of the thesis, a new frame skipping transcoder is proposed to greatly reduce the computational complexity and reduce the quality degradation. The proposed architecture is mainly performed on the discrete cosine transform (DCT)

domain to achieve a low complexity transcoder. It is observed that the re-encoding error is significantly reduced at the frame-skipping transcoder when the strategy of a direct summation of DCT coefficients is employed. By using the proposed frame-skipping transcoder, the video qualities of the active sub-sequences can be improved significantly.

Besides, most video conferencing systems use DCT-based encoders. However, under low bit rates, a DCT-based encoder exhibits visually annoying blocking artifacts. Recently, wavelets have been used in internet applications. The major advantage of using a wavelet is its high quality and the absence of blocking artifacts when compared to the conventional video encoder. Although a wavelet-based coder can achieve a good quality, its computational speed is an area of concern. Motivated by this, a new region-based video coder architecture is proposed to achieve a good video quality with a low complexity. The proposed video coder is based on the adaptive region-based updating technique by which the video is updated according to the motion activity. A simple and fast object tracking technique is proposed to locate the region of interest. Features of the proposal includes (i) a user-specified region of interest selection as to which the region can be changed by the user at any time instance and (ii)an adaptive bit allocation that allows the user to specify the relative quality between the foreground and the background to increase the interactivity. This architecture guarantees a high video quality in the region of interest while reducing the overall bit rate and the computation time even under low bit rates.

In the second half of the thesis, we address a problem of speech enhancement, which is to recover a speech source from a mixture of its delayed versions and additive noise. By using the constrained optimisation technique, an algorithm based on the second

order statistics is developed. The new proposed algorithm requires no strong limitations to the speech signal and the noise. Simulation results show that our algorithm achieves a better performance as compared to other algorithms.

Finally, although the MPEG Audio provides the perceptual lossless audio compression, the demanded bitrate and the computational complexity are higher than the conventional speech coding approach. Motivated by this, a fast bit allocation algorithm for the MPEG audio encoder is proposed, which is able to generate an identical MPEG bitstream produced by the standard bit allocation algorithm described in MPEG audio standard. The proposed algorithm employs the bit allocation information of the previous frame as a reference for allocating the restricted bits to each of the 32 subbands in the current frame such that the number of iterations can be significantly reduced. Results of the study show that the performance of the proposed bit allocation algorithm works well at different encoded bitrates.

It is exciting to report in this thesis that significant gains in terms of computation and scalability can be achieved by employing our adaptive approaches. Undoubtedly, these adaptive techniques can enable the video conferencing to become more scaleable and provide good quality video and audio in practical situations.

Acknowledgements

I would like to express my greatest thank to my Supervisor, Professor W.C. Siu, for his continuous encouragement, guidance and care during the period that I worked on this thesis. He spent much of his invaluable time with me discussing my research, reviewing and correcting my articles and drafts of this thesis. His profound knowledge and experience in digital signal processing, his rigorous approach to reserach , many interesting and stimulating lectures that he has given, and his hard working style and devotion to science and reaserch, have inspired me during my work on the thesis. This will continually influence my future research and career.

I would like to thank all colleagues in the Mutlimedia Signal Processing Centre for their friendship, support and encouragement, especially, Dr. Christie Chan, Dr. Y.L. Chan, Mr. W.F Cheung, Mr. K.P. Cheung, Mr. W.L. Hui, Dr. Kenneth Lam, Dr. Bonnie Law, Dr. C.T.Leung, Dr. Daniel Lun, Mr. K.W. Wong and Mr. W.H. Wong.

It is my pleasure to acknowledge the Research Degrees Committee of the Hong Kong Polytechnic University for its generous support over the years.

Above all, I thank my family for their constant love, encouragement and support. Without their understanding and patience, it is impossible for me to complete this research study.

Statement of Originality

The following contributions reported in this thesis are claimed to be original.

Low-Complexity and High-Quality Frame-Skipping Transcoder for Continuous Presence Multipoint Video Conferencing

1. Direct summation of DCT coefficients for macroblock without motion compensation has been Proposed. (Chapter 3, section 3.3.1)

   The proposed architecture is mainly performed on the discrete cosine transform (DCT) domain to achieve a low complexity transcoder. It is observed that the re-encoding error is avoided at the frame-skipping transcoder when the strategy of direct summation of DCT coefficients is employed.

2. DCT-domain buffer updating for motion-compensated macroblock. (Chapter 3, section 3.3.2)

   In order to reduce the implementational complexity of the motion-compensated macroblock, a cache subsystem is added to our proposed transcoder. Since motion compensation of multiple macroblocks may require the same pixel data, a cache subsystem is implemented to reduce redundant inverse quantization, inverse DCT and motion compensation computations. The arrangement is significant since the frequency of caching hits is high. This is due to the fact that the locality of motion is often present within each frame.

3. Indirect summation of DCT coefficents for motion-compensated macroblock. (Chapter 3, section 3.3.2)

   It is observed that our proposed approach on indirect summation of DCT coefficents for motion-compensated macroblock has introduced less error degradation as compared to the conventional appraoch.

4. Multiple Frame-skipping transcoding approach for video combiner in multipoint video conferencing. (Chapter 3, section 3.3.3)

   When multiple frames are dropped, it can be processed in the forward order, thus eliminating the multiple DCT-domain buffers that are needed to store the incoming quantized DCT coefficients of all dropped frames. Furthermore, the proposed frame-skipping transcoder can be used to realize the continuous presence multipoint video conferencing. By using the proposed frame-skipping transcoder and dynamically allocating more frames to the active participants in video combining, we are able to achieve uniform PSNR performance of the sub-sequences and improve significantly the video qualities of the active sub-sequences.

Proposed an Architecture for a Region-based object tracking video coder for Multipoint Video Conferencing using wavelet transform.

5. The main features of our proposed wavelet-based video coder include: 1) a user-specified region of interest selection as to which the region can be changed by the user at any I-frame; 2) a dynamic region tracking technique by which the video is tracked and updated according to the motion activity and 3) an adaptive bit allocation that allows the user to specify the relative quality between the foreground and the background. (Chapter 4)

   This architecture guarantees a high video quality in the region of interest while reducing the overall bit rate and the computation time. Experimental results confirm that the approach produces a good video quality even under low bit rates.

A Fast Bit Allocation Algorithm for MPEG Audio Encoder

6. A fast bit allocation algorithm for the MPEG audio encoder is proposed. (Chapter 5, section 5.1)

   The proposed algorithm uses the bit allocation information of the previous audio frame as a reference for allocating the restricted bits to each of the 32 subbands in the current audio frame such that the number of iterations can be significantly reduced. A process of bit reallocation is also suggested to ensure the generation of an identical MPEG bitstream produced by the standard bit allocation algorithm described in the MPEG audio standard. The result shows that the speed-up of the proposed algorithm is remarkable at different encoded bitrates.

A Constrained Optimisation Approach to Speech Signal Recovery

7. A Constrained Optimisation Approach to Speech Signal Recovery. (Chapter 5, section 5.2)

   We address a problem of speech enhancement, which is to recover a speech source from a mixture of its delayed versions and additive noise. By using the constrained optimisation technique, the second order statistics based algorithm is developed. The new proposed algorithm makes no strong assumptions on the speech signal and the type of noise. Simulation results show that our algorithm achieves a better performance as compared to other algorithms.

Table of Contents

List of Figures

List of Tables

Author's Publications

(List of Publications of the Author on which this thesis is based)

International Journal Papers

Recent Submissions

1. K.T. Fung, Y.L.Chan and W.C. Siu, "Low-Complexity and High-Quality Frame-Skipping Transcoder for Continuous Presence Multipoint Video Conferencing" submitted to IEEE Transactions on Multimedia.

2. K.T. Fung, Y.L.Chan and W.C. Siu, "A Fast Bit Allocation Algorithm for MPEG Audio Encoder", submitted to Electronics letters.

International Conference Papers

Paper Published or Provisionally Accepted

1. W.Li, K.T.Fung and W.C.Siu "A Constrained Optimisation Approach to Speech Signal Recovery," Proceedings, 1999 International Symposium on Signal Processing and Intelligent System (ISSPIS'99), pp.400-403, November, 1999.

2. K.T.Fung, Y.L.Chan and W.C.Siu "A Fast Bit Allocation Algorithm for MPEG Audio Encoder", Proceedings, 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing (ISIMP'2001), pp.5-8, May, 2001.

3. K. T. Fung, Y. L. Chan and W. C. Siu, "Low-Complexity and High Quality Frame-Skipping Transcoder," Proceedings, IEEE International Symposium on Circuits and Systems (ISCAS'2001), pp.29-32, May, 2001.

4. K. T. Fung, N. F. Law and W. C. Siu, "Region-based object tracking for Multipoint Video Conferencing using wavelet transform," Proceedings, IEEE International Conference on Consumer Electronic(ICCE'2001) , pp.268-269, June, 2001

Chapter 1

Introduction and Motivation

1.1 Generic video coding in video conferencing

In a video conferencing system, it usually consists of a numbers of units, such as the input video source, input audio source, video encoder, communication channel, video decoder, audio encoder, audio decoder, output display unit and speaker. Figure 1.1 shows the architecture a video conferencing system for software implementation. In the video encoding process, the camera captures the image raw data and passes it into the video encoder for video compression. Typical video encoder is block-based, the nature which will be discussed in detail in the next chapter. In the audio encoding process, the microphone receives the audio signal and passes it into the audio encoder for audio compression and the detail of the audio encoder will be discussed in the next chapter as well. The outputs of the video and audio encoder are bitstreams. Then these bitstreams are transmitted to the other conference participant for decoding. In the decoding process, the decoder looks at the compressed bitstreams and passes them to the video and audio decoders. In order to have a good performance of a video conferencing, the video and audio synchronization is particularly important to the present study.

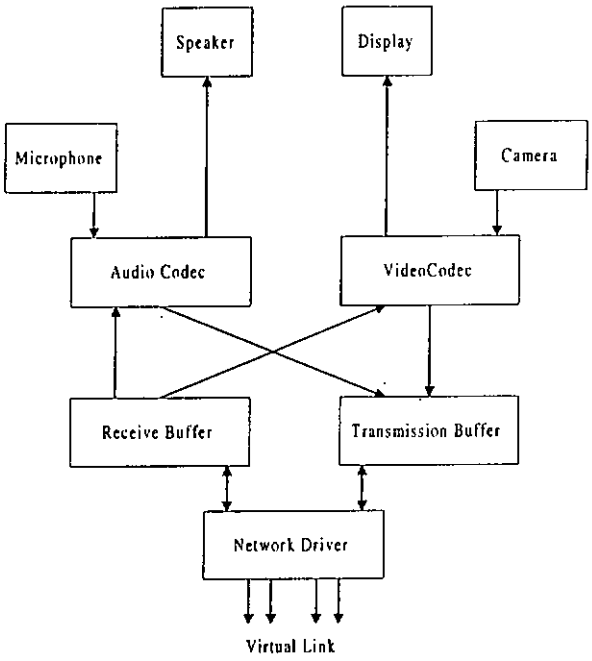


Figure 1.1. Architecture of the video conferencing system for software implementation.

The architecture of our multipoint video conferencing system for software implementation is shown in Figure 1.1. Figure 1.2 shows the details of a front video Encoder, a video transcoder and an end decoder in the multipoint video conferencing system[1]. The function of the video transcoder is used to convert a previously compressed bit stream into a lower bitrate bit stream and will be discuss in details in chapter 2. First, the captured image in the client side is passed into the front encoder. In the front encoder, the image data undergo the 2D-DCT transform[2] in order to achieve energy compaction. Since our human eyes are less sensitive to the high frequency, so the high frequency components will be discard in the quantization process to achieve high compression ratio. Then the quantized DCT coefficients are performed the variable length coding which encodes the DCT coefficients according to the statistical behaviours. The variable length coding tries to use less bits to represent the DCT coefficients which appear more frequently in order to improve coding efficiency. However, only use the

spatial information is not enough to achieve high compression ratio. So, some motion estimation and compensation algorithms are used to reduce the temporary redundancy[3-5]. These algorithms search for the best match for the current frame from a previous frame to reconstruct the motion compensated frame. The motion estimation is performed on the luminance marcoblocks based on the sum of absolute difference(SAD) or mean square error(MSE). Due to it simplicity, SAD is widely use as a measure criterion. In order to obtain a motion vector for the current macroblock, the best matching block that results in a minimal SAD is searched within a predefined search area S in the previous reconstructed reference frame such that

$$(u_t^S, v_t^S) = \arg \min_{(m,n)\in S} SAD_S(m,n)$$

where

$$SAD_S(m,n) = \sum\sum |P_S^c(i,j) - R_S^p(i+m, j+n)|$$

and m and n are the horizontal and vertical components of the motion vector. The $P_S^c(i,j)$ and $R_S^p(i+m, j+n)$ represents a pixel in the current frame and a displaced pixel by (m,n) in the previous reconstructed reference frame respectively. The superscript "c" or "p" denotes the "current" or "previous" frame respectively, and the subscript "f" or "s" indicates the "front" or "second" encoder. However, the motion compensated frame has a significant video quality degradation. In order to improve the video quality, the prediction errors (the difference between the motion compensated frame and the previous frame) are also transmitted in order to improve the video quality. In this case, the difference frame, not the original image, is encoded using the DCT. In the end-decoder, the received bitstream is undergo the inverse of the variable length coding to get the

quantized DCT coefficients. Then the quantized DCT coefficients are arranged for inverse quantization and inverse DCT to reconstruct the encoded image. If the motion estimation and motion compensation algorithm are used, the decoder needs a buffer to store the reference frame and perform motion compensation.

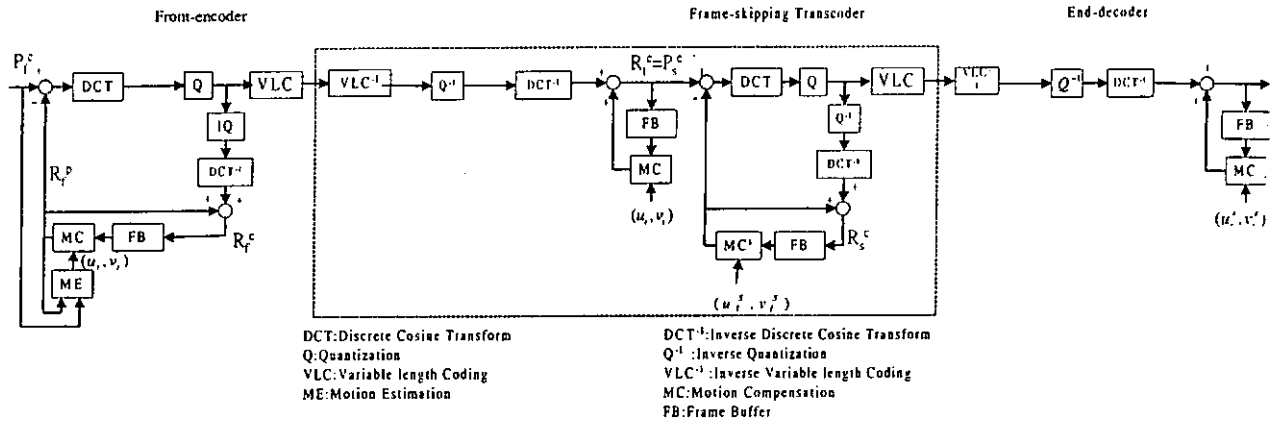


Figure 1.2 The detail of the front video Encoder, video transcoder and end decoder in the multipoint video conferencing system.

1.2 Video transcoding

In this section, we give a brief introduction to a few common video conferencing systems reported in the literature. These video conferencing systems usually have a Multipoint control unit (MCU) to perform the video transcoding mentioned in the previous section and its function is to convert a previously compressed bit stream into a lower bitrate bit stream. These techniques can usually be characterized by the coded-domain transcoding[6] and pixel-domain transcoding[7]. The coded-domain transcoding is mainly performed in the coded domain and mainly gives focus on the header manipulation such as to modify the header from four QCIF format to CIF format. In this approach, a QCIF-to-CIF combiner combines four H.261 bitstream coded in CIF picture format. This method uses the property that each H.261 group of blocks(GOB) data has a

synchronization code word so that an H.261 bitstream can be parsed into units of GOB data without performing any decoding operation[8]. The computational complexity is the lowest and the incurred delay is less than the pixel-domain transcoding since no re-encoding process is performed, however, the flexibility is the lowest since the this approach requires an asymmetric network channel between a user terminal and the bridge because the video bit rate from the bridge to the terminals is four times that from the terminals to the bridge for four conference participants involved[9]. This requirement is not supported by most networks and codecs. Another approach is the pixel-domain transcoding which is mainly performed in the pixel domain. The input bitsteam needs to decode into pixel domain and the decoded video signal is re-encoded at the desired output bit rate. So, the bit rate conversion can be achieved to support difference network environments. However, it involves high computational complexity, memory and delay since it includes the double re-encoding process. Recently, the DCT-domain transcoding is emerging, the operations of which are performed in the DCT domain[10]. This approach decodes an input video bitstream partially to the discrete cosine transform(DCT) coefficient level, and the decoded DCT coefficients are requantized with a larger quantization step size than that originally used at the encoding terminal. The computational speed is much faster than the pixel-domain transcoding since no fully decoding process or re-encoding process is needed. Also, this approach supports bit rate conversion as well. However, the problem in this approach is that the requantization error will accumulate and the prediction memory mismatch at the decoder will cause poor video quality. Motivated by this, the DCT domain transcoding performs in the spatial domain using the direct requantization and feedback technique to avoid error propagation

is proposed recently[11]. And some re-using techniques such as motion vector refinement[1,12], motion vector interpolation[13] are suggested to further improve the computational complexity. Due to its low-complexity and high flexibility, the DCT-domain transcoding becomes widely used in video transcoding. Besides the spatial domain transcoding, some researchers give focus on video transcoding which reduces the temporal redundancy by using a frame-skipping technique[13]. Since in most video conferencing, only one or two conference participants are active at the same time. This techniques become more useful since a larger transcoding ratio can be achieved while maintain the video quality. A detailed analysis of video transcoding will be discussed in the Chapter 2.

1.3 Motivation and research objectives

In this research, we have made use of the above architecture to implement the multipoint video conferencing system. There are two major issues that we are concerned with. They are the video and audio qualities in video conferencing system.

For the video processing part, we consider multipoint video conferencing over a wide-area network or a video server. The system has to support quality of service to heterogeneous clients or transmission channels. It is in this scenario that the video server has to have the capability to perform transcoding which is regarded as a process of converting a previously compressed video bitstream into a lower bitrate bitstream without modifying its original structure. One straightforward approach for implementing a transcoding is to cascade a decoder and an encoder[1,12-14], commonly known as pixel-domain transcoding. The incoming video bitstream is decoded in the pixel domain, and

the decoded video frame is re-encoded at the desired output bitrate according to capability of the clients' devices and available bandwidth of network. This involves high processing complexity, memory, and delay. As a consequence, some information reusing approaches[1,12] have been proposed. For example, motion vectors extracted from the incoming bitstream after the decoding can be used to significantly reduce the complexity of the transcoding. Besides, the video quality of the pixel-domain transcoding approach suffers from intrinsic double-encoding process, which introduces additional degradation. This problem will be discussed in detail in chapter 2. One of the objectives in this thesis is to decide a low complexity and high quality of video transcoder which will be discuss in the chapter 3.

With the advance of video compression and networking technologies, multipoint video conferencing becomes popular in the consumer market[15]. Most video conferencing systems use DCT-based encoders. A good performance can be achieved with a large bandwidth[16]. However, under low bit rates, the DCT-based encoder exhibits visually annoying blocking artifacts. Recently, wavelets have been used in internet applications. The major advantage of using a wavelet is its high quality and the absence of blocking artifacts when compared to the conventional video encoder[17-18]. Although a wavelet-based coder can achieve a good quality, its computational speed is an area of concern. A way to speed up the computation is to explore the fact that the various regions in an image are not of equal importance. This concept has been adopted in dynamic bit allocation and frame-skipping technique. In this thesis, a new region-based video coder architecture is proposed to achieve a good video quality with a low complexity. The proposed video coder is based on the adaptive region-based updating

technique by which the video is updated according to the motion activity. This architecture allows a high quality video in the region of interest while reducing the overall bit rate and computation time. Since the user of a video conferencing system might be in fast motion when active, a simple and fast object tracking technique is proposed to locate the region of interest. This approach produces a good video quality even under low bit rates. Since the video encoder is wavelet-based, blocking artifacts are avoided. Our proposed region-based video encoder has two major features: 1)selection of the region of interest and 2) adaptive bit allocation for the foreground (region of interest) and the background. The purpose of region selection is to identify the region of interest in an image, e.g. the speaker's face in video conferencing. This region is updated automatically by tracking the object's motion. The wavelet-based coder is then applied separately to the foreground and background. Because the size of the region of interest is small, the computation time could be reduced significantly. Adaptive bit allocation is then performed. It makes sure that the video quality of the foreground is always better than that of the background. This is particularly important for unstable networks or low bit rate applications.

In the audio processing, noises are very irritating and sometimes seriously degrade conversation quality. It is, therefore, desirable to cancel or at least significantly reduce such noise to enhance speech conversation quality[19]. We have tried to address the problem of speech enhancement, which is to recover a speech source from a mixture of its delayed versions and additive noise. By using the constrained optimisation technique, an algorithm based on the second order statistics is developed. The new proposed algorithm requires no strong limitations to the speech signal and the noise, which is more practical

in real life situation. Experimental results show that our algorithm achieves a better performance as compared to other algorithms.

Although in conventional video conferencing system we adopt the speech codec to encode the speech signal, MPEG Audio Coding may be a good choice for speech compression or voice mail application to achieve high quality of audio compression. MPEG Audio provides the perceptual lossless audio compression, however, the demanded bitrate and the computational complexity are higher than the conventional speech coding approach[20-22]. In this thesis, a fast bit allocation algorithm for the MPEG audio encoder is proposed, which is able to generate an identical MPEG bitstream produced by the standard bit allocation algorithm described in MPEG audio standard. The proposed algorithm employs the bit allocation information of the previous frame as a reference for allocating the restricted bits to each of the 32 subbands in the current frame such that the number of iterations can be significantly reduced. Results of the study show that the performance of the proposed bit allocation algorithm works well at different encoded bitrates.

1.4 Organization of the thesis

Before embarking on a description of the main topic of research, a review of current multipoint video conferencing techniques is given in Chapter 2. This review gives focus on practical aspects of a practical video conferencing system such as transcoding in video, video and audio synchronization, blind signal separation and audio encoder.
For the sake of the easy identification, we may consider that the following chapters are divided into two parts and described as below.

Chapter 3 and 4 form part I of the thesis. Both chapters mainly discuss techniques on a scalable video coding. The emphasis is on adaptive schemes which take active speaker into account for bit allocation. In Chapter 3, a new frame skipping transcoder is proposed to greatly reduce the computational complexity and reduce the quality degradation. The proposed architecture is mainly performed on the discrete cosine transform (DCT) domain to achieve a low complexity transcoder. It is observed that the re-encoding error is significantly reduced at the frame-skipping transcoder when the strategy of a direct summation of DCT coefficients is employed. By using the proposed frame-skipping transcoder, the video qualities of the active sub-sequences can be improved significantly. In Chapter 4, a new region-based video coder architecture is proposed to achieve a good video quality with a low complexity. The proposed video coder is based on the adaptive region-based updating technique by which the video is updated according to the motion activity. A simple and fast object tracking technique is proposed to locate the region of interest. Features of the proposal includes (i) a user-specified region of interest selection as to which the region can be changed by the user at any time instance and (ii) an adaptive bit allocation that allows the user to specify the relative quality between the foreground and the background to increase the interactivity. This architecture guarantees a high video quality in the region of interest while reducing the overall bit rate and the computation time even under low bit rates.

Chapter 5 forms part II of the thesis. This chapter mainly discusses the techniques on the audio processing. Emphasis is on adaptive schemes which reduce computational complexity and enhance speech quality. In the first half of this chapter, we address a problem of speech enhancement, which is to recover a speech source from a mixture of

its delayed versions and additive noise. By using the constrained optimization technique, an algorithm based on the second order statistics is developed. The new proposed algorithm requires no strong limitations to the speech signal and the noise. Simulation results show that our algorithm achieves a better performance as compared to other algorithms. In the second half of this chapter, a fast bit allocation algorithm for the MPEG audio encoder is proposed, which is able to generate an identical MPEG bitstream produced by the standard bit allocation algorithm described in MPEG audio standard. The proposed algorithm employs the bit allocation information of the previous frame as a reference for allocating the restricted bits to each of the 32 subbands in the current frame such that the number of iterations can be significantly reduced. Results of the study show that the performance of the proposed bit allocation algorithm works well at different encoded bitrates.

Chapter 6 is devoted to a summary of the work herein and the conclusions reached as a result. Suggestions are also included for further research in this area and in the general area of video conferencing system.

Chapter 2

Review of Current Techniques in Video Conferencing

2.1Multipoint Video Conferencing

In this chapter, we will have more detailed discussion on multipoint video conferencing and a review of a number of current techniques in video conferencing will be given. With the advance of video compression, networking technologies and international standards, video conferencing is widely used in our daily life[3,8,23-27]. In recent years, more and more video conferencing products are appearing in the market. The rapid growth of video conferencing has driven the development of multipoint video conferencing that can be integrated into personal computers for providing an efficient way for more than two people to exchange information at multiple locations.

For multipoint video conferencing over a wide-area network, the conference participants are connected to a multipoint control unit (MCU) [28-29] which coordinates and distributes audio, video and data streams among multiple participants in multipoint video conferencing according to the requirement of channel bandwidth. Figure 2.1 shows the scenario of four persons participating in multipoint video conferencing with a MCU. An audio mixer in the MCU accepts audio data in a variety of formats, with different data rates. The audio mixer must be required to decode and mix these different audio bitstreams from all conference participants and the mixed audio signal is encoded again for distributing to the conference participants. Similarly, a video combiner is also included in the MCU to combine the multiple coded video bitstreams from the conference participants into a coded video bitstream which conforms to the video coding

standard such as H.263[3], and sends it back to the conference participants for decoding and presentation.

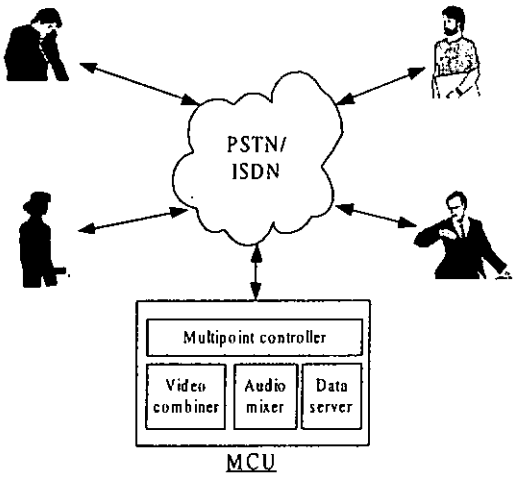


Figure 2.1 An example of multipoint video conferencing.

The four-point video conferencing as shown in Figure 2.1 is over practical wide-area network such as Public Switch Telephone Network (PSTN) or Integrated Service Digital Network (ISDN) in which the channel bandwidth is constant and symmetrical. Assuming the encoded video bitstream of each conference participant is R kb/s in a quarter common intermediate format (QCIF: 176 × 144 pixels). The MCU receives and decodes the multiple video bitstreams from all conference participants. The decoded videos are combined in a common intermediate format (CIF: 352 × 288 pixels) through the video combiner. An example of the video combining of four QCIF frames into a single CIF frame is illustrated in Figure2.2. The combined video is re-encoded at R kb/s in order to fulfill the requirement of channel bandwidth for sending back the encoded video to all conference participants. Therefore video transcoding is required to perform at the video combiner. In the previous chapter, we have been introduced about video transcoding. In this chapter, a detailed analysis of video transcoding will be given. Figure 2.3 shows the block diagram of video combining for multipoint video

conferencing by the transcoding approach[10,14]. This approach is pixel-domain transcoding. The QCIF bitstream at rate R of a conference participant firstly passes into the decoder as shown in figure 2.3. Then the image in pixel domain is stored in frame buffer. The pixel-domain multiplexer combines the pixel-domain frame buffer in CIF format and passes it into the encoder to produce the bitstream at rate R (for symmetric network) or lower rate for difference applications. We refer this approach to as the conventional approach used in a video conferencing system. In the next section, some recent techniques of transcoding will be discussed in details.

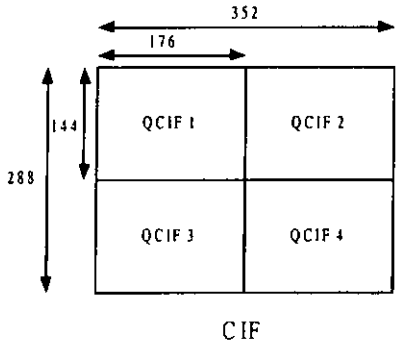


Figure 2.2. Combining four QCIF frames into a single CIF frame.

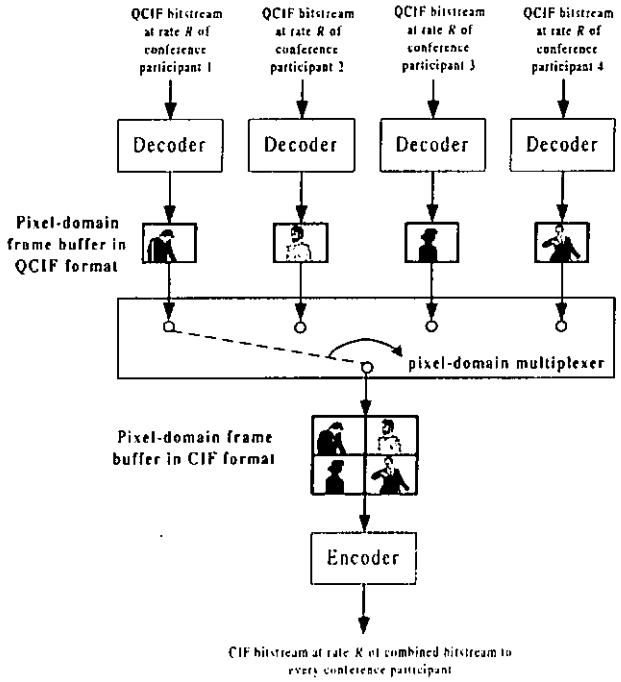


Figure 2.3. Pixel-domain video combiner using transcoding approach.

2.2Multipoint Video Transcoding techniques

Transcoding is a very practical approach for video combining in multipoint video conferencing over a symmetrical wide-area network. However, the computational complexity is inevitably increased since the individual video bitstream needs to be decoded and the combined video signal needs to be encoded as described in previous section. In addition, the video quality of the transcoding approach suffers from its intrinsic double-encoding process which introduces additional degradation and will be described in detail in the following. The visual quality and the computational complexity need to be considered in video transcoding of multipoint video conferencing. In order to provide satisfactory visual quality of combined and transcoded video, the re-distribution of limited bits in the reencoding process to different parts of combined video will be critical. In most multipoint video conferencing, usually only one or two conference participants are active and talking at any given time, while the other participants are listening with little motion. To make the best use of the available bit rates, a rate control scheme is proposed in [7] to measure the motion activity of each sub-sequence by computing the sum of magnitudes of its corresponding motion vectors, and allocate the bit rates to each sub-sequence according to its activity. Consequently, more bits will be allocated to those sub-sequences with higher motion activities, and this control scheme will produce much more uniform visual quality. Since more bits need to represent the image with higher motion activity in practice, this approach of bit re-allocation process becomes more dynamic. Although more uniform video can be achieved, however, by only re-allocate bits to the moving region in spatial domain is not enough to provide good quality for the active speaker in a practical situation. In general, we always focus on an

active speaker instead of inactives speakers. It will be better if an active speaker has good video quality instead of the inactive ones.

In recent years, the Discrete Cosine Transform (DCT) domain transcoding was introduced [10,30-31], under which the incoming video bitstream is partially decoded to form the DCT coefficients and downscaled by the requantization of the DCT coefficients. Since the DCT-domain transcoding is carried out in the coded domain where complete decoding and re-encoding are not required, the processing complexity is significantly reduced. The problem, however, with this approach is that the quantization error will accumulate, and prediction memory mismatch at the decoder will cause poor video quality. This phenomenon is called "drift" degradation, which often results in an unacceptable video quality. Thus, several techniques for eliminating the "drift" degradation [10,30-31] have been proposed. The DCT-domain transcoding is a very attractive approach for many video applications. However, it is impossible to achieve the desired output bitrate by performing only the requantization. In other words, if the bandwidth of the outgoing channel is not enough to allocate bits with requantization, frame skipping is a good strategy for controlling the bitrate and maintaining the picture quality within an acceptable level. It is difficult to perform frame skipping in the DCT-domain since the prediction error of each frame is computed from its immediate past frames. This means that the incoming quantized DCT coefficients of the residual signal are no longer valid because they refer to the frames which have been dropped. This problem has not been fully considered in the literature. In the following, we would like to describe this problem in details. Figure 2.4 shows the structure of a conventional frame-skipping transcoder in pixel-domain [1,6-7]. At the front encoder, the motion

vector, mv_t, for a marcoblock with the size of $N \times N$ in the current frame is computed [32-36] by searching the best matched macroblock within a search window S in the previous reconstructed frame as shown in Figure 2.5 and it is obtained as follows:

$$mv_t = (u_t, v_t) = \arg \min_{(m,n) \in S} SAD(m,n) \tag{2.1}$$

$$SAD(m,n) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} | O_t(i,j) - R_{t-1}(i+m, j+n) | \tag{2.2}$$

and m and n are the horizontal and vertical components of the displacement of a matching macroblock, $O_t(i,j)$ and $R_{t-1}(i,j)$ represent a pixel in the current frame t and in the previous reconstructed reference frame t-1, respectively.

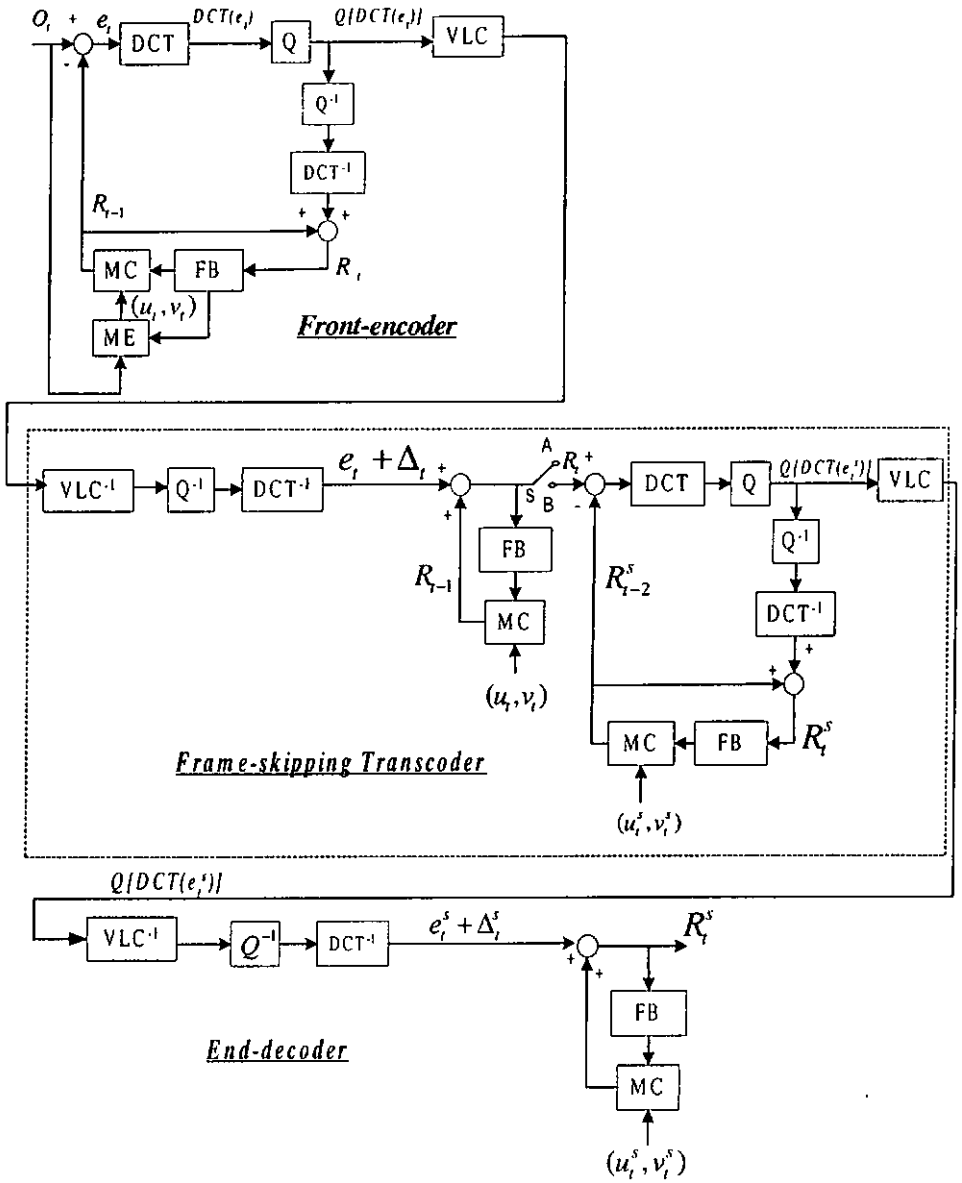


Figure 2.4 The structure of a conventional frame-skipping transcoder in pixel-domain.

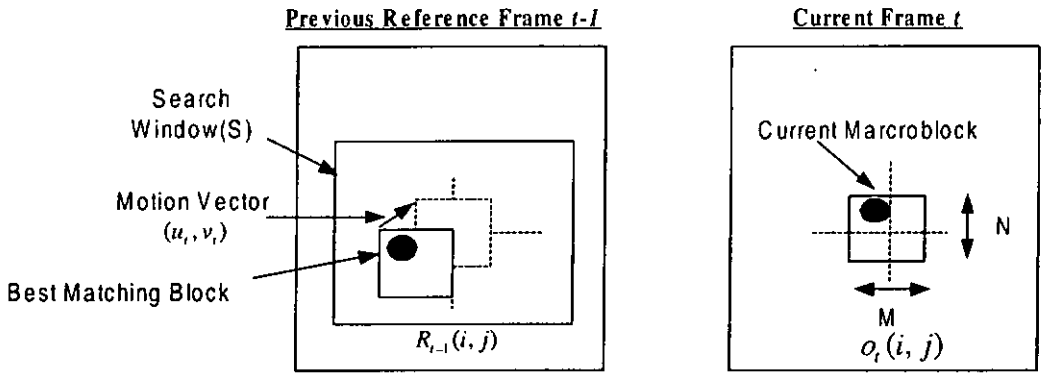


Figure 2.5 shows Block matching motion estimation.

In transcoding the compressed video bitstream, the output bitrate is lower than the input bitrate. As a result, the outgoing frame rate in the transcoder by cascading a decoder and an encoder is usually much lower than the incoming frame rate. Hence switch *S* is used to control the desired frame rate of the transcoder. Table 2.1 summaries the operating modes of the frame-skipping transcoder.

Table 2.1 Switch position for different modes of frame skipping.

Frame skipping mode	*S* Position
Skipped frame	*A*
Non-skipped frame	*B*

Assume that frame *t-1*, R_{t-1}, is skipped. However, R_{t-1} is required to act as the reference frame for the reconstruction of frame *t*, R_t, such that

$$R_t(i, j) = R_{t-1}(i + u_t, j + v_t) + e_t(i, j) + \Delta_t(i, j) \tag{2.3}$$

where $\Delta_t(i, j)$ represents the reconstruction error of the current frame in the front-encoder due to the quantization, and $e_t(i, j)$ is the residual signal between the current frame and the motion-compensated frame,

$$e_t(i, j) = O_t(i, j) - R_{t-1}(i + u_t, j + v_t) \tag{2.4}$$

Substituting (2.4) into (2.3), we obtain the expression for R_t,

$$R_t(i,j) = O_t(i,j) + \Delta_t(i,j) \tag{2.5}$$

In the transcoder, an optimized motion vector for the outgoing bitstream can be obtained by applying the motion estimation such that

$$mv_t^S = (u_t^S, v_t^S) = \arg \min_{(m,n)\in S} SAD^S(m,n) \tag{2.6}$$

$$SAD^S(m,n) = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} | R_t(i,j) - R_{t-2}^S(i+m, j+n) | \tag{2.7}$$

where $R_{t-2}^s(i,j)$ denotes a reconstructed pixel in the previous non-skipped reference frame. The superscript "s" is used to denote the symbol after performing the frame-skipping transcoder. Although the optimized motion vector can be obtained by a new motion estimation, it is not desirable because of its high computational complexity. Reuse of the incoming motion vectors has been widely accepted because it is considered to be almost as good as performing a new full-scale motion estimation and was assumed in many transcoder architectures[1,12]. Thus, we assume that the new motion vector be (u_t^s, v_t^s). Hence, the reconstructed pixel in the current frame after the end-decoder is,

$$R_t^s(i,j) = R_{t-2}^s(i+u_t^s, j+v_t^s) + e_t^s(i,j) + \Delta_t^s(i,j) \tag{2.8}$$

where $e_t^s(i,j) = R_t(i,j) - R_{t-2}^s(i+u_t^s, j+v_t^s)$ and $\Delta_t^s(i,j)$ represents the requanization error due to the re-encoding in the transcoder, then,

$$R_t^s(i,j) = O_t(i,j) + \Delta_t(i,j) + \Delta_t^s(i,j) \tag{2.9}$$

This equation implies that the reconstructed quality of the non-skipped frame deviates from the input sequence to the transcoder, R_t. Re-encoding of the current frame involves a re-computation of the residual signal between the current frame and the non-skipped reference frame. Note that frame *t-2* acts the reference instead of frame *t-1*, since the

frame *t-1* does not exist after frame skipping. The newly quantized DCT-domain data are then re-computed by means of the DCT and quantization processes. This re-encoding procedure can lead to an additional error Δ_t'. The effect of the re-encoding error is depicted in Figure 2.6 where the "Salesman" sequence was transcoded at half of the incoming frame-rate. In the figure, the peak signal-to-noise ratio (PSNR) of the frame-skipping pictures is plotted to compare with that of the same pictures directly using a decoder without a transcoder. This figure shows that the re-encoding error leads to a drop in picture quality of about 3.5dB on average, which is a significant degradation.

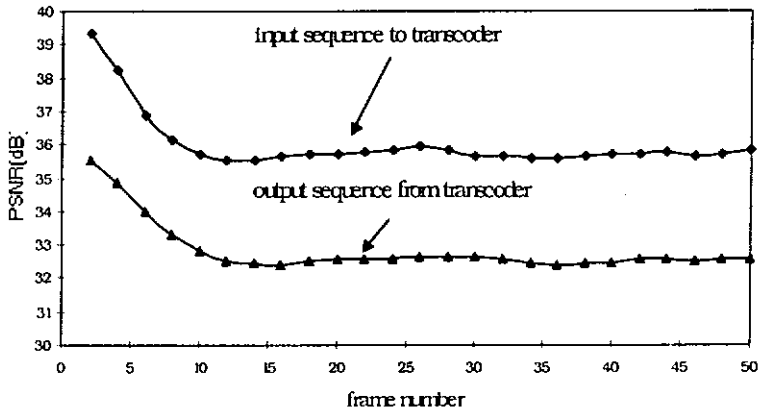


Figure 2.6. Quality degradation of conventional frame-skipping transcoder for the "Salesman" sequence.

Besides the quality degradation, the pixel-domain transcoder also has a high processing complexity. This is due to the fact that the skipped frame must be decompressed completely, and should act as the reference frame to the non-skipped frame for reconstruction.

2.3 MPEG Encoder Skeleton

In this section, one of the famous video coding scheme-MPEG(Moving Pictures Expert Group) will be discussed. This video coding scheme is a compression standard

for storage of moving images on storage media such as compact disc, digital audio tape, hard disc and the features in MPEG can random access, shuttle search with visible image, slow motion and freeze frame. More importantly, simple and cheap decoder relatively to encoder makes MPEG become more and more popular in multimedia application.

Before having a detailed analysis about the MPEG video encoder, the objective of MPEG standard will be given. In 1991, the MPEG 1 standard is emerged and its bitrate is about 1.5Mbit/s and a typical image sizes is 288x352pixels, without interlace and the frame rate is about 24 to 30. This standard is widely used nowadays(e.g. VCD). In 1994, the MPEG 2 standard inherited the MPEG 1 and provides flexible chroma format(4:4:4, 4:2:2, 4:2:0) and support interlace. It also give focus on broadcasting applications. The image quality includes at least NTSC, PAL at 3-5Mbit/s, and HDTV at 20Mbit/s. In 1998, the MPEG 4 standard inherited the MPEG 2 and focus on interactive functionalities. New functions such as content-based multimedia data access, manipulation and bitstream editing, hybrid coding of natural and synthetic AV objects are provided. In 2001, the MPEG 7 standard is introduced with a focus on the content representation standard for information search, storage and retrieval.

In this thesis, a MPEG Video Encoder will be described in detail. This video encoder is very similar to the video transcoder which is widely used in the video conferencing system[11,37]. Figure 2.7 shows the general architecture of a MPEG Video Encoder.

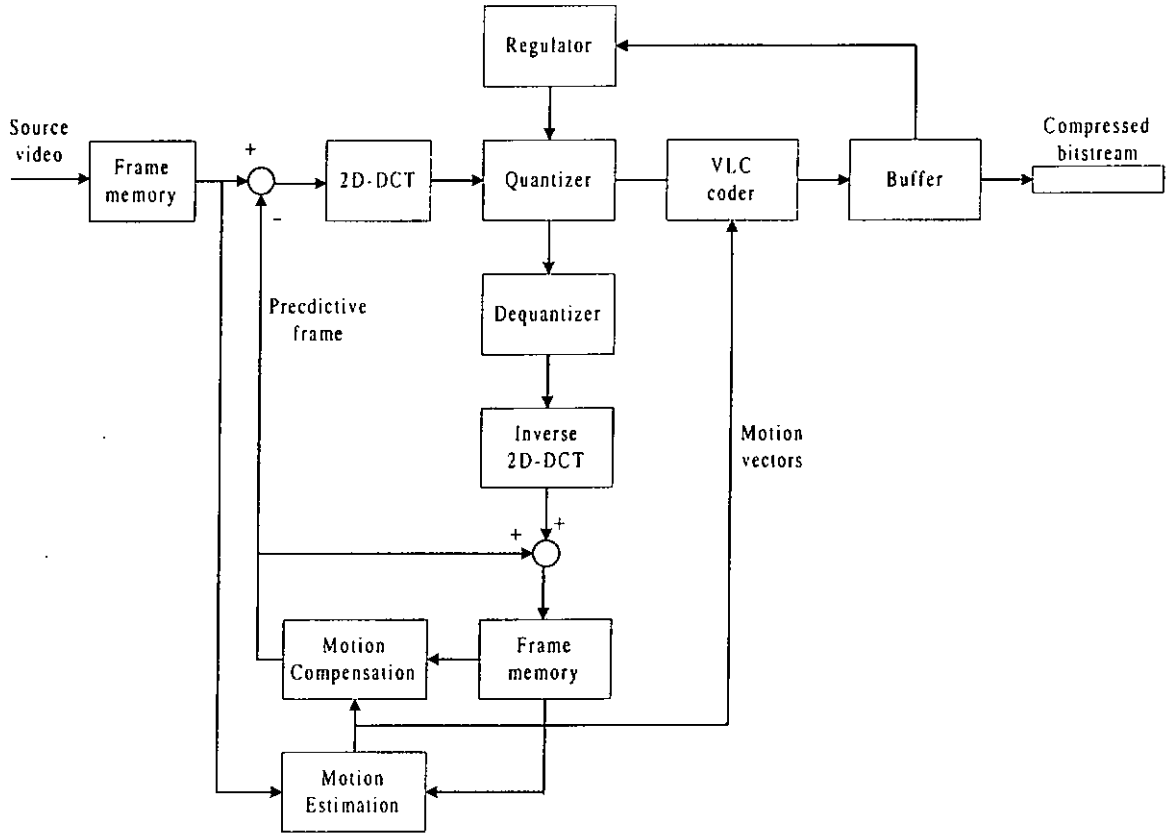


Figure 2.7 shows the block diagram of MPEG Video Encoder.

The video source is firstly passed into the memory and then the image is divided into macroblocks with the size 16x16 and each macroblock contains four blocks with the size of 8x8 pixels(see figure 2.8). For Intraframe(I-frame) compression, no motion estimation and motion compensation are needed. Figure 2.9 shows a block diagram for I-frame compression. Each block of the input frame is coded with 2D-FDCT for energy compaction. Most of the high frequency component will be discard in the quantizer. After the quantization process, the DCT coefficients will perform a zig-zag scanning. Then the variable length coder(VLC) performs a entropy coding by using some tables provided in [4-5] to form a bitstream.

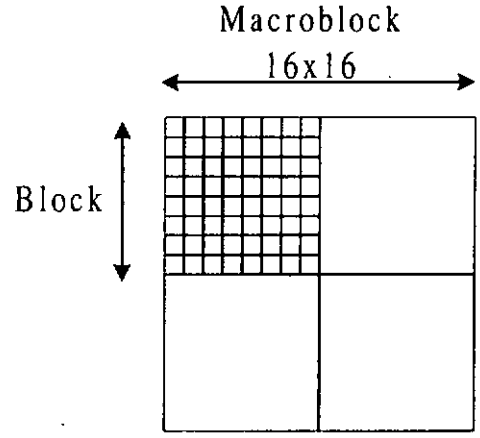


Figure 2.8 shows the structure of marcoblock and block.

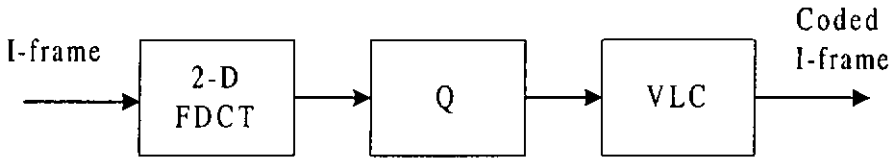


Figure 2.9 shows the block diagram for I-frame compression.

For the coding of a predicted frame(P-frame), motion estimation and motion compensation is employed. Figure 2.10 shows the P-frame compression. Frame 1 is the reference frame and frame 5 is the current frame, say for example the motion estimator looks at each macroblock in frame 5 and tries to find out the best match from reference frame 1 and this process is known as motion estimation. The motion vector for a macroblock in the current frame is computed by searching all possible locations within a predefined search window S in the previous reconstructed reference frame as shown in Figure 2.5. The motion vector is defined as the displacement of the best matching block from the position of the current macroblock and it is obtained as follows:

$$(u_t, v_t) = \arg \min_{(m,n) \in S} SAD(m,n) \tag{2.10}$$

$$SAD(m,n) = \sum_{i}^{M} \sum_{j}^{N} |O_t(i,j) - R_{t-1}(i+m, j+n)| \tag{2.11}$$

where m and n are the horizontal and vertical components of the displacement of a matching macroblock, $O_t(i, j)$ and $R_{t-1}(i, j)$ represent a pixel in the current frame at time t and in the previous reconstructed reference frame at time t-1, respectively.

After the motion estimation process, the motion vectors are obtained. By using these motion vectors and the reference frame 1, we can reconstruct the predicted frame. However, the reconstructed frame quality is degraded significantly as compared to the original current frame. In order to improve the video quality of this reconstructed frame, the difference between the original current frame and this reconstructed frame(prediction error) is encoded. This prediction error is coded using the 2D-DCT, quantization and VLC and the rest of the process is similar to the I-frame compression.

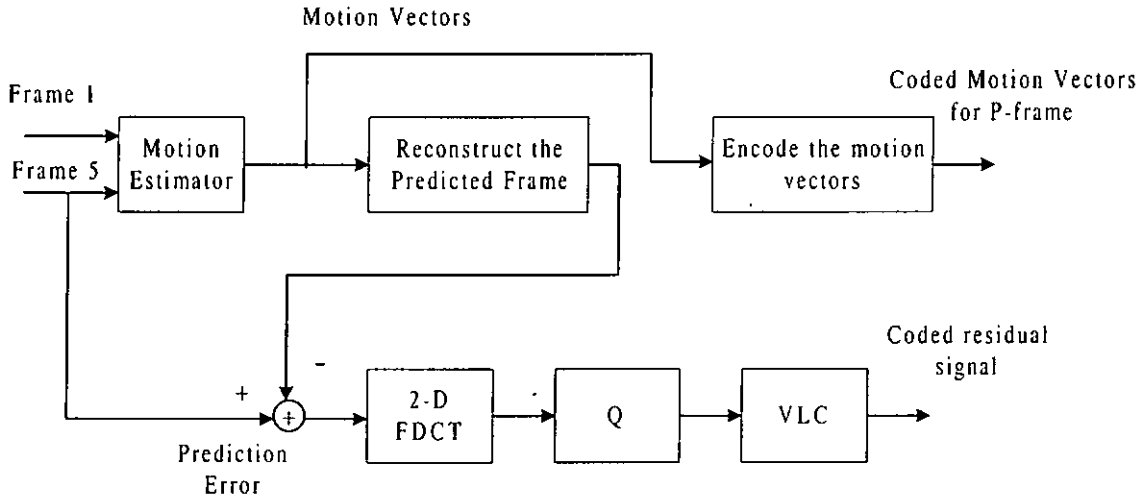


Figure 2.10 shows the P-frame compression.

2.4Audio and Video Synchronization

It will be annoying to the conference participants if the speech signal does not match the video signal. This phenomenon occurs more frequently when encoding delay is large or the network is unstable. Figure 2.11 describe how the video and audio

information can be synchronized with each other[38]. Firstly, the system time clock (STC) is used to record the time information in the client side. Then the captured video and the audio frames are record by the presentation time stamp(PTS). After the encoding process of the video and audio, the system clock reference (SCR) is used to record the time for the encoded bitrstream before transmission. Then end-to-end synchronization can be performs in the receiving sides. By comparing the PTS and SCR, the decoded video and audio can be presented without jittering. And this technique will become more important when frame-skipping transcoder is employed. Figure 2.12 illustrates the idea of video and audio synchronization. The video frame1, frame2 and frame3 will be presented at 0.25 sec, 0.4sec and 0.55sec respectively instead of present them directly. Less jittering can be observed due to the arrangement that the time difference of frame 2 to frame 3 is reduced from 0.2 second to 0.15sec in this approach. This approach can help us to play the video more smoothly and to do video synchronization. Also, Audio synchronization has to be at the same time[38].

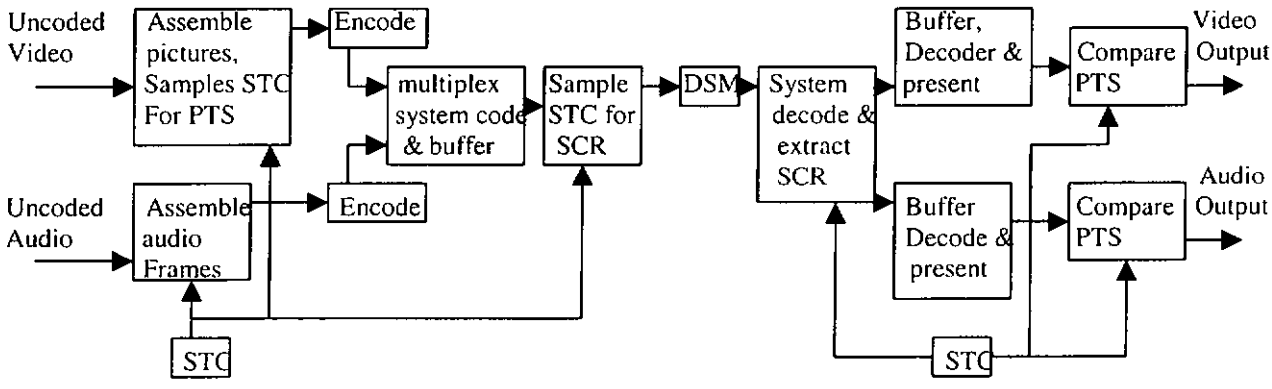


Figure 2.11 shows the block diagram of synchronization of video and audio.

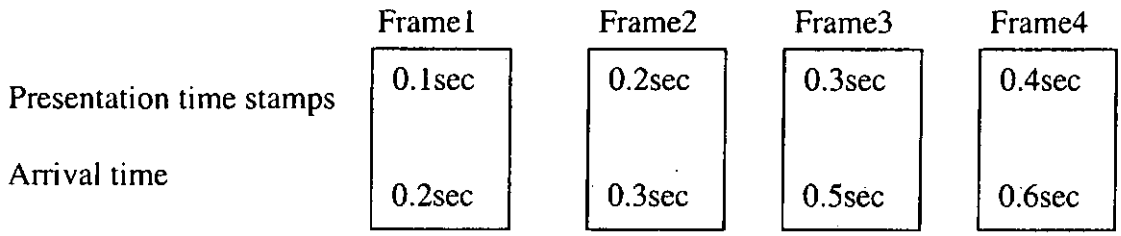


Figure 2.12 Illustration of video and audio synchronization.

2.5 Blind signal separation

Speech enhancement has been active in speech signal processing. The objective is to extract a single speech source signal from its delayed versions and in a noisy environment. Depending upon the amount and the type of noise, and the strength of echoes existing in the environment, the resulting speech signals could vary substantially. The quality of the speech may range from being slightly degraded to being annoying to listeners, and in the worst case it could be totally unintelligible. It is very necessary to recover the speech signal from the distortion.

A number of approaches to signal recovery have been proposed [39-42]. These approaches make use of the output second-order statistics [39,40] or the output higher-order statistics [41,42]. They are basically the least squares solutions. However, if the unknown parameters in an algorithm have inherent non-linear relations, they are usually assumed to be independent from each other in order to apply the linear least squares technique. A larger estimation error is inevitably generated, although additional post-processing step may reduce the error.

In order to illustrate how does the blind signal separation extract one or more source signals from a set of measurements, a simple model using output decorrelation

will be presented. Equations 2.12 and 2.13 show the equations of modeling the received signals.

$$x_1(t) = s_1(t) + as_2(t) \tag{2.12}$$

$$x_2(t) = bs_1(t) + s_2(t) \tag{2.13}$$

where $s_1(t)$ and $s_2(t)$ represent two difference input sources at time t. $x_1(t)$ and $x_2(t)$ represent the received signal at time t in receiver 1 and receiver 2 respectively, and a and b are arbitrary constants. Figure 2.13 shows the model of the received signal and Figure 2.14 shows the model of the output signal.

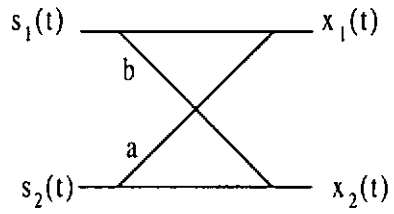


Figure 2.13 Modeling the received signals

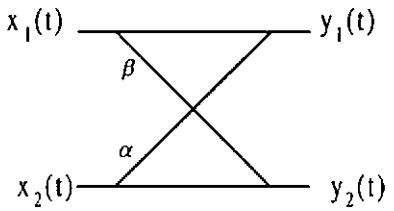


Figure 2.14 Modeling the output signal

$y_1(t)$ and $y_2(t)$ represent the output signals where α and β are the arbitrary constants. Equations 2.14 and 2.15 describe the relationship between the received signals and the output signals. Then the estimated signal can be derived by equations 2.16 and 2.17.

$$y_1(t) = x_1(t) + \alpha\, x_2(t) \tag{2.14}$$

$$y_2(t) = \beta\, x_1(t) + x_2(t) \tag{2.15}$$

$$s_1(t) = \frac{1}{1-\beta\alpha} y_1(t)$$
$$= \frac{1}{1-\beta\alpha}[(1+\alpha b)s_1(t) + (a+\alpha)s_2(t)] \quad (2.16)$$

$$s_2(t) = \frac{1}{1-\beta\alpha} y_2(t)$$

$$= \frac{1}{1-\beta\alpha}[(b+\beta)s_1(t) + (1+a\beta)s_2(t)] \quad (2.17)$$

The cross-correlation function is:

$$r_{y1y2}(q) = \beta\, r_{x1x2}(q) + \alpha\, r_{x2x2}(q) + (1+\alpha\beta) r_{x1x2}(q) \quad (2.18)$$

When the signal is separated at the output, then $r_{y1y2}(q)=0$ is the condition to be satisfied.

$$\text{i.e. } \beta\, r_{x1x2} + \alpha\, r_{x2x2} + (1+\beta\alpha) r_{x1x2} = 0 \quad (2.19)$$

By dividing equation (2.19) by $(1+\alpha\beta)$ for $\alpha\beta \neq 1$, the equation becomes:

$$\frac{\beta}{1+\alpha\beta} r_{x1x1} + \frac{\alpha}{1+\alpha\beta} r_{x2x2} = -r_{x1x2} \quad (2.20)$$

By using the least square principle, the optimal values for α and β from this over-determined set of equation($q \geq 2$) can be estimated.

For $R=[r_{x1x1}\ r_{x2x2}]$, $p=-r_{x1x2}$

$$w^T = [\alpha_t \quad \beta_t]$$

$$\text{where } \alpha_t = \frac{\alpha}{1+\alpha\beta} \quad \beta_t = \frac{\beta}{1+\alpha\beta}$$

$$\therefore \frac{\alpha_t}{\beta_t} = \frac{\alpha}{\beta} \quad (2.21)$$

$$\alpha = \alpha_t + \alpha\alpha_t\beta \Rightarrow \alpha = \frac{\alpha_t}{(1-\alpha_t\beta)}$$

$$\beta = \beta_t + \alpha\beta\beta_t \Rightarrow \beta = \frac{\beta_t + \alpha_t\beta\beta_t}{(1-\alpha_t\beta)}$$

$$\beta - \alpha_t\beta^2 = \beta_t - \alpha_t\beta\beta_t + \alpha_t\beta\beta_t$$

$$\Rightarrow \alpha_t\beta^2 + \beta_t - \beta$$

By solving β from the above quadratic equation and calculate the parameter α by the relationship in the equation (2.21).

Hence, by substituting α and the corresponding β into the following equations:

$$y_1(t)=x_1(t)+ \alpha x_2(t)$$

$$y_2(t)=x_2(t)+ \beta x_1(t)$$

The estimated signals can be obtained as shown below by divided the above equation by $\frac{1}{1-\alpha\beta}$:

$$s_1(t) = \frac{1}{1-\beta\alpha} y_1(t)$$

$$s_2(t) = \frac{1}{1-\beta\alpha} y_2(t)$$

By using this approach, the estimated signals are very similar to the original input signals in theoretical simulation. However, the estimation cannot be guaranteed completely due to the nature of the sources in a practical environment. It may be due to the assumption of the signals need to be stationary and the constants a and b actually vary quite significantly with time.

2.6 Audio Encoder

In the previous section, the MPEG video coding scheme is introduced. In this section, we would like to describe the MPEG audio coding scheme. The MPEG audio

coding scheme[43] is a perceptual audio coding standard and consists of three audio coding algorithms called layer I, II and III. Among these three layers, layer I and layer II have been widely used in multimedia and broadcasting related products[44,45]. Which layer is employed for an application of audio coding is determined by the computational complexity and performance required by the application[46].

Without loss of generality, layer II audio encoder will be discussed in this section and its basic structure is shown in Figure 2.15. The input audio samples first pass through a subband filter which divides the input signal into 32 equal-width frequency subbands as shown in figure 2.16. These subband samples are then uniformly quantized according to the bit allocator, which decides the quantization manners with consideration to the audio quality and the required bits. In aid of the bit allocator, the psychoacoustic model provides the perceptual resolution, which dynamically calculates the masking threshold for each subband. Any audio signals with levels falling below the corresponding masking thresholds are imperceptible to human ear. Then, the objective of the bit allocator is used to dynamically distribute the available bits among the subbands according to the masking threshold provided by the psychoacoustic model. Theoretically, by coding at the demanded bitrate, a perceptually lossless quality of audio signals can be achieved. If the demanded bitrate is higher than the available bitrate, the bit allocator tries to minimize the total noise-to-mask ratio(NMR) over the frame with the constraint that the number of bits required does not exceed the number of bits available for the frame. After the bit allocation process, the subband samples are then scaled, quantized according to the bit allocation information, and formatted into an encoded MPEG bitstream together with a header, bit allocation and scaling information. In the standard

bit allocation procedure, over 100 iterations in average are needed which is computational intensive.

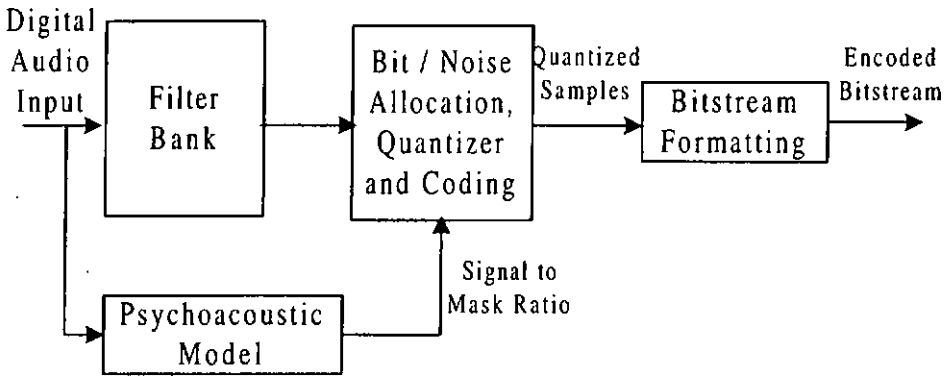


Figure 2.15. Basic structure of the audio encoder.

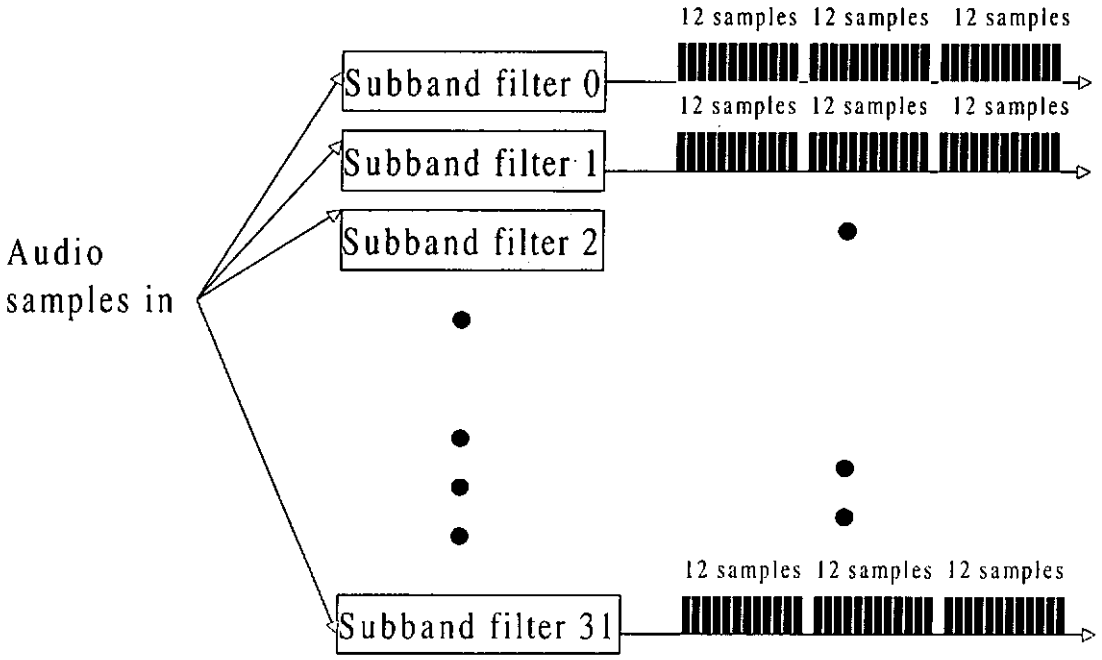


Figure 2.16 Subband filtering

The MPEG coding scheme achieves the compression by placing the quantization noise in the frequency subbands where the ear is least sensitive. The psychoacoustic model determines from the maximum noise level of the input audio which would just be perceptible (masking level) for each of the subbands as shown in figure 2.17. The quantitative sketch of figure 2.18 gives a few more details about the masking threshold.

Within a critical band, tones below this threshold are masked. As the amount of the quantization noise is directly related to the number of bits used by the quantizer, the bit allocation algorithm assigns the available bits in a manner which minimizes the audible distortion. The psychoacoustic model described in the standard returns the Signal to Mask Ratio (SMR) for each subband, which is defined as the difference in dB between the level of masker and the minimum masking threshold within the critical band. Assuming an m-bit quantizaion of an audio signal, within the critical band the quantization noise will not be audible as long as its signal-to-noise ratio(SNR) is higher than its SMR. The bit allocation algorithm computes the Noise-to-Mask Ratio(NMR) from the SMR using the following expression.

$$NMR(m)=SMR-SNR(m) \quad \text{(in dB)} \tag{2.22}$$

NMR(m) describes the difference in dB between the SMR and the SNR to be expected from an m-bit quantization. The NMR value is also the difference (in dB) between the level of quantization noise and the level where a distortion may just become audible in a given subband. Within a critical band, coding noise will not be audible as long as NMR(m) is negative.

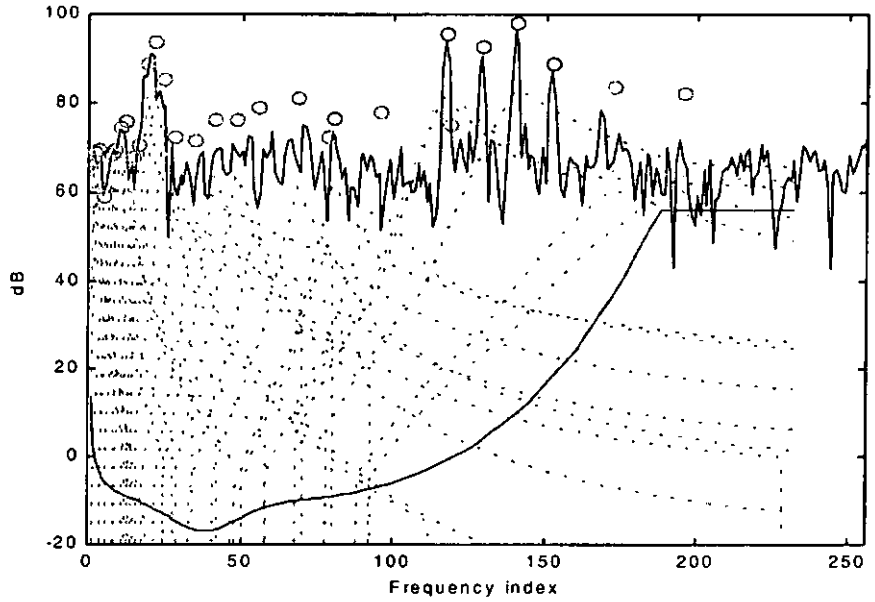


Figure 2.17 Calculation of the masking threshold

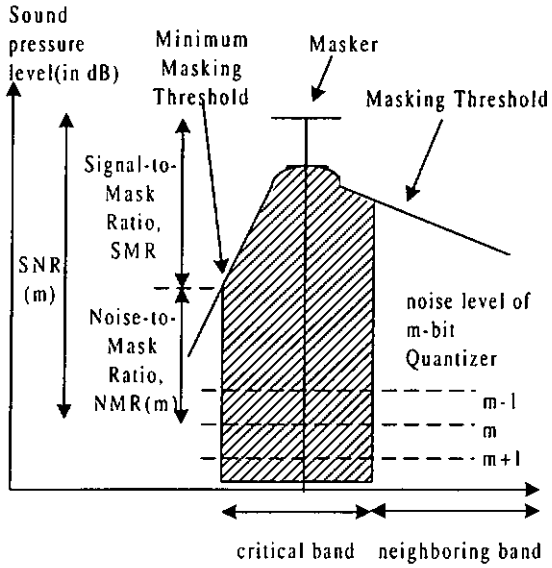
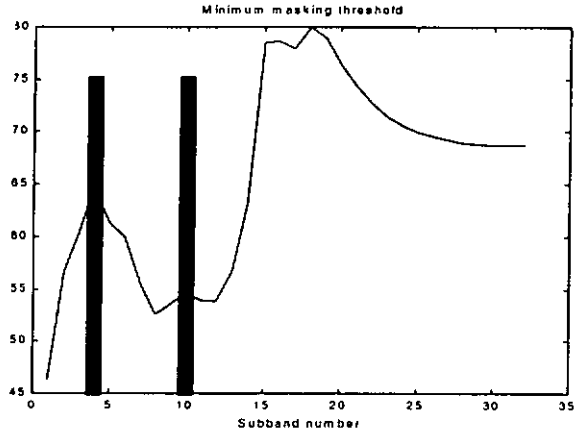


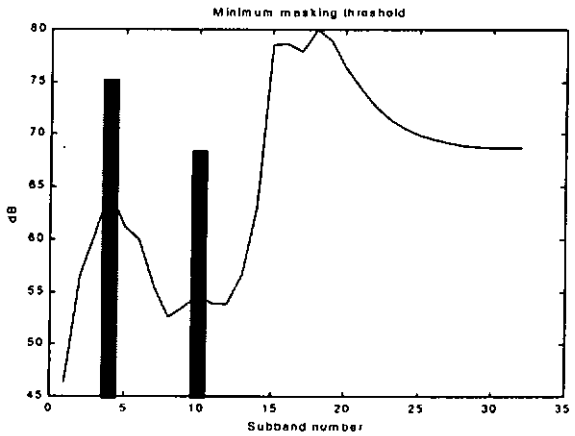
Figure 2.18 Masking threshold and signal-to-mask ratio(SMR).

The bit allocator looks at both the outputs samples from the filterbank and the SMR from the psychoacoustic model, and adjusts the bit allocation in order simultaneously to meet both the bitrate requirements and the masking requirements. Bit

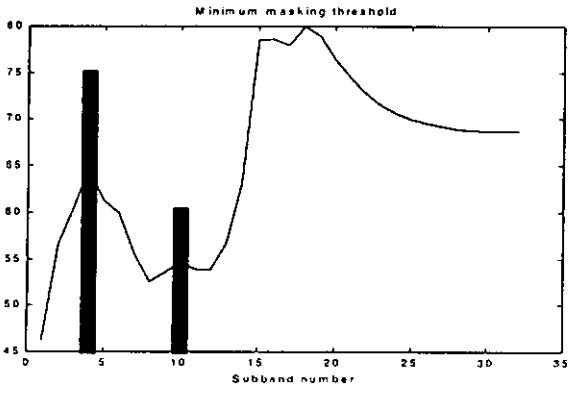
allocation algorithm recommended by the MPEG standard involves a number of iterations where, in each iteration, the number of quantizing levels of the subband with the largest NMR is decreased as long as the number of bits used does not exceed the total number of bits available for the frame. The NMR in dB for each subband is obtained from equation (2.22). Each incremental bit assigned to a subband will incur 36 bits as there are a total of 36 time samples within each subband. As a result, the iteration will stop once the number of bits available for coding is smaller than 36 as shown in figure 2.19.



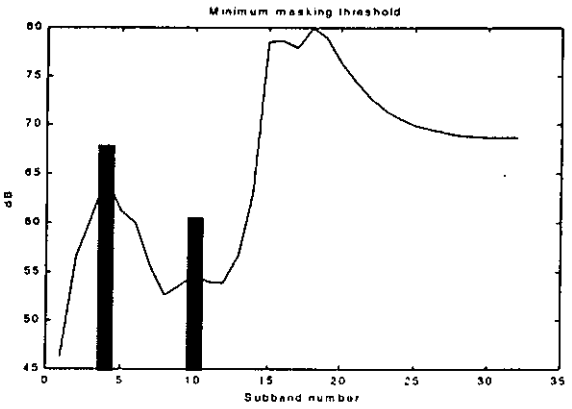
(a)



(b)



(c)



(d)

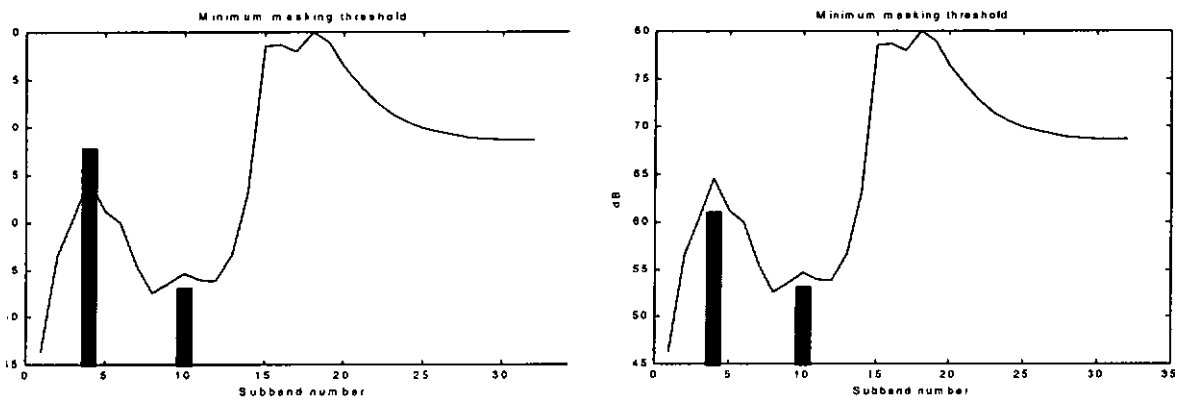


Figure 2.19 The iteration process of bit allocation in conventional approach.

Without loss of generality, the scheme with 2 subbands as a comparison instead of 32 subbands as a comparison is used for easy illustration. By comparing the NMR in these two subbands, one more bit will be assigned to the subbands with the smaller NMR in order to minimize the most noticeable quantization noise. From figure 2.19(a-f), the iteration will continue to allow more bits to the subband which has a larger noticeable quantization noise if bits are available. If NMR is negative among the 32 subbands, perceptually lossless compression can be achieved.

In practice, the procedure of allocating the bits involves a large number of iterations. Therefore, it demands an enormous amount of computational steps. Recently, a fast bit allocation algorithm has been proposed[47], which seeks to reduce the number of the iterations by computing the demand bitrates, which is defined as the bit allocation needed to give a zero or just negative value of NMR for every subband. i.e instead of allocate bits to the subband with the largest NMR in the conventional approach, all subband is first assign bits in every subband to make the NMR become zero or just negative. This demand bitrate is then substracted from the total available bit rate as constrained by the channel to obtain the available bitrate for allocation. If the available

bitrate is greater than or equal to 36 bits per frame, the available bits will be allocated according to the standard iterative procedure until all the bits are exhausted. However, if the available bit rate is less than zero, the subband with the smallest NMR is identified and the number of quantizing steps allocated to it is reduced and, consequently, the NMR of the subband is also increased. This iterative procedure is carried out until the available bitrate becomes positive. The number of iterations of this algorithm is significantly reduced when the allowable bitrate is very close to the demanded bitrate. However, it breaks down at very low bitrates because of the large number of iterations needed to reduce the number of bits allocated to each band.

Chapter 3

Low-Complexity and High-Quality Frame-Skipping Transcoder for Continuous Presence Multipoint Video Conferencing

3.1 Introduction

This chapter presents a new frame-skipping transcoding approach for video combiner in multipoint video conferencing. Combining the multiple compressed video sequences from conference participants into a single video sequence would result in a total bitrate which might overwhelm the outgoing channel bandwidth. Transcoding is regarded as a process of converting a previously compressed video bitstream into a lower bitrate bitstream. The high transcoding ratio may result in an unacceptable picture quality when the incoming video bitstream is transcoded with the full frame rate. Frame skipping is often used as an efficient scheme to allocate more bits to the representative frames, so that an acceptable quality for each frame can be maintained. However, the skipped frame must be decompressed completely, and should act as the reference frame to the non-skipped frame for reconstruction. The newly quantized DCT coefficients of prediction error need to be re-computed for the non-skipped frame with reference to the previous non-skipped frame; this can create an undesirable complexity in the real time application as well as introduce re-encoding error. A new frame-skipping transcoding architecture for improved picture quality and reduced complexity is proposed. The proposed architecture is mainly performed on the discrete cosine transform (DCT) domain to achieve a low complexity transcoder. It is observed that the re-encoding error is avoided at the frame-skipping transcoder when the strategy of direct summation of DCT coefficients is employed. Furthermore, the proposed frame-skipping transcoder can be used to realize the continuous presence multipoint video conferencing. We observe

that, in most multipoint video conferencing, usually only one or two participants are active at any given time, while the other participants are listening with little motion. To achieve similar video quality for all the conference participants, the active participants require higher frame rates than the inactive participants. By using the proposed frame-skipping transcoder and dynamically allocating more frames to the active participants in video combining, we are able to make more uniform PSNR performance of the sub-sequences and the video qualities of the active sub-sequences can be improved significantly.

With the advance of video compression and networking technologies, multipoint video conferencing is becoming more and more popular [3-9,48-50]. In multipoint video conferencing, the conference participants are connected to a multipoint control unit (MCU) which receives video signals from several different participants, processes the received video signals and transmits the processed video signals to all participants. Multipoint video conferencing can be, for example, the "switched presence" type or the "continuous presence" type. A typical switched presence MCU [28-29] permits the selection of a particular video signal from one participant for transmission to all participants. Switched presence MCU generally does not require the processing of video signals to generate a combined video signal and therefore is relatively simple to implement. Switched presence MCU is not ideal for multipoint video conferencing applications since only one participant can be seen at a given time. A better choice is one based on the continuous presence mode [6-7,9,48]. Continuous presence MCU consists of a video combiner which combines the multiple coded video bitstreams from the conference participants into a single coded video bitstream and sends it back to the

conference participants for decoding and presentation. Each participant in a continuous presence conference can then view one or more of the other participants in real time.

There are two possible approaches to implement a video combiner for continuous presence multipoint video conferencing. The first approach is coded-domain combining [6,9] which modifies the headers of the individual coded bitstreams from the conference participants, multiplexes bitstreams, and generates new headers to produce a combined video bitstream conforming to the video coding standard. For example, a QCIF-to-CIF combiner is proposed in [6] which concatenates four H.261 bitstreams coded in QCIF picture format (176 × 144 pixels) into a single H.261 bitstream coded in CIF picture format (352 × 288 pixels). Since the coded-domain combiner only needs to perform the multiplexing and header-modification functions in concatenating the video bitstreams, the implementation complexity is very low. Also, since it does not need to decode and re-encode the video sequence, it does not introduce any quality degradation. However, the coded-domain combiner requires an asymmetric network channel between the participant and the MCU because the video bitrate from the MCU to the participants is four times that from the participants to the MCU. This asymmetric requirement is not supported by most networks.

The second approach to video combining is based on the transcoding technique [7,48]. This type of video combiner decodes each coded video bitstream, combines the decoded video in the pixel domain, and re-encodes the combined video at the transmission channel rate. Transcoding is a very practical approach for video combining in multipoint video conferencing over a symmetrical wide-area network. However, the computational complexity is inevitably increased since the individual video bitstream

needs to be decoded and the combined video signal needs to be encoded. As a consequence, some information reusing approaches [1,12] have been proposed, in which certain information, such as motion vectors, that have been extracted from the incoming bitstream after decoding can be used to significantly reduce the complexity of the video combiner. The video quality of the transcoding approach suffers from its intrinsic double-encoding process since the video from each participant needs to be decoded, combined, and then re-encoded, which introduces additional degradation.

In recent years, the Discrete Cosine Transform (DCT) domain transcoding was introduced [10,30-31], under which the incoming video bitstream is partially decoded to form the DCT coefficients and downscaled by the requantization of the DCT coefficients. Since the DCT-domain transcoding is carried out in the coded domain where complete decoding and re-encoding are not required, the processing complexity is significantly reduced. The problem, however, with this approach is that the quantization error will accumulate, and prediction memory mismatch at the decoder will cause poor video quality. This phenomenon is called "drift" degradation, which often results in an unacceptable video quality. Thus, several techniques for eliminating the "drift" degradation [10,30-31] have been proposed. The DCT-domain approach is a very attractive approach for video combining in multipoint video conferencing. The continuous presence architecture on the DCT-domain lies between the full decoding/encoding pixel-domain approach and the coded-domain combiner approach. The asymmetric network channel is not required due to the requantization of the DCT coefficients However, it is impossible to achieve the desired output bitrate by performing only the requanization. In other words, if the bandwidth of the outgoing channel is not

enough to allocate bits with requantization, frame skipping is a good strategy for controlling the bitrate and maintaining the picture quality within an acceptable level. It is difficult to perform frame skipping in the DCT-domain since the prediction error of each frame is computed from its immediate past frames. This means that the incoming quantized DCT coefficients of the residual signal are no longer valid because they refer to the frames which have been dropped. Although frame-rate reduction in the transcoder is useful for multipoint video conferencing, only a few papers have been published in the literature on this topic [1,12-13]. In this chapter, we provide a computationally efficient solution to perform frame skipping in a transcoder, mainly in the DCT-domain, to avoid the complexity arising from pixel-domain transcoding. A new system architecture for continuous presence multipoint videoconferencing based on the proposed low-complexity and high-quality frame-skipping transcoder is developed. Simulation results are presented to show the performance improvement realised by our proposed architecture.

The organization of this chapter is as follows. Section 3.2 presents an in-depth study of the re-encoding error in the frame-skipping transcoder. The proposed frame-skipping transcoder is then described in Section 3.3. Section 3.4 presents the system architecture of the proposed continuous presence in a multipoint videoconference. Simulation results are presented in Section 3.5. Finally, some conclusive remarks are provided in Section 3.6.

3.2 Frame-skipping Transcoding

In recent years, several international standards such as H.261[8], H.263[3], MPEG1[4] and MPEG2[5] have been established to support various video coding applications. All of these standards employ techniques for exploiting two types of

redundancies in the uncompressed video signal to achieve the desired compression gain. First, preserving only significant DCT coefficients can considerably eliminate the spatial redundancy between pixels within a single frame because of the energy compaction property of the DCT. Furthermore, the motion-compensated predictive coding scheme is used to remove the temporal redundancy between frames. Motion compensation is defined as the process of compensating for the displacement of moving objects from one frame to another. In practice, motion compensation is preceded by motion estimation[32-36] which is the process of finding motion vectors among frames. A motion-compensated block in the previous reconstructed reference frame is then subtracted from the current macroblock and this residual signal is encoded by using DCT to further remove the spatial redundancy.

The motion vector for a macroblock in the current frame is computed by searching all possible locations within a predefined search window S in the previous reconstructed reference frame as shown in Figure 3.1. The motion vector is defined as the displacement of the best matching block from the position of the current macroblock and it is obtained as follows:

$$(u_t, v_t) = \arg \min_{(m,n) \in S} SAD\,(m,n) \tag{3.1}$$

$$SAD(m,n) = \sum_{i}^{M} \sum_{j}^{N} | O_t(i,j) - R_{t-1}(i+m, j+n) | \tag{3.2}$$

where m and n are the horizontal and vertical components of the displacement of a matching macroblock, $O_t(i,j)$ and $R_{t-1}(i,j)$ represent a pixel in the current frame t and in the previous reconstructed reference frame t-1, respectively.

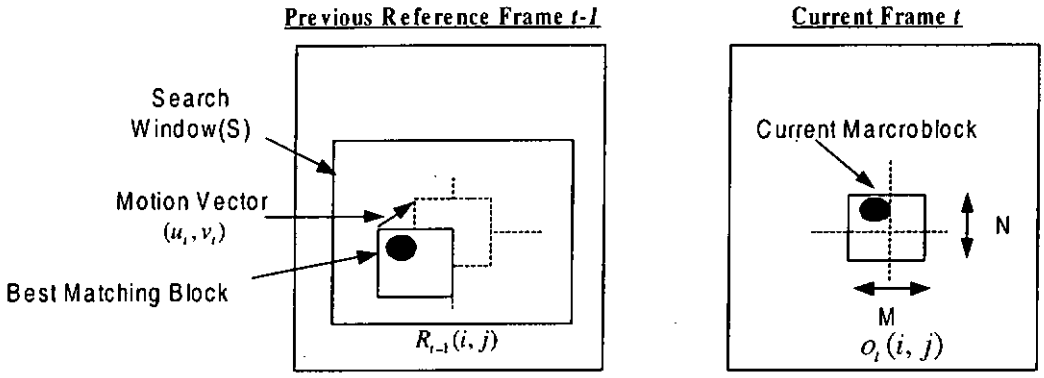


Figure 3.1. Block matching motion estimation.

Although the requantization of DCT coefficients provides a simple and fast transcoder [10,30-31], it may not be enough to achieve the desired output bitrate by performing only the requanization with an acceptable quality. If frame skipping is performed in the transcoder, the control of the output bitrate becomes more flexible. However, the skipped frame must be decompressed completely, and should act as the reference frame to the non-skipped frame for reconstruction. All motion vectors and predicted errors must be computed again for the non-skipped frame which references the previous non-skipped frame. This process can create undesirable complexity in real time applications.

Figure 3.2 shows the structure of a conventional frame-skipping transcoder in pixel-domain [1,12,14]. Since the output bitrate is lower than the input bitrate, the outgoing frame rate in the transcoder by cascading a decoder and an encoder is usually much lower than the incoming frame rate. Hence switch *S* is used to control the desired frame rate of the transcoder. Table 3.1 summaries the operating modes of the frame-skipping transcoder.

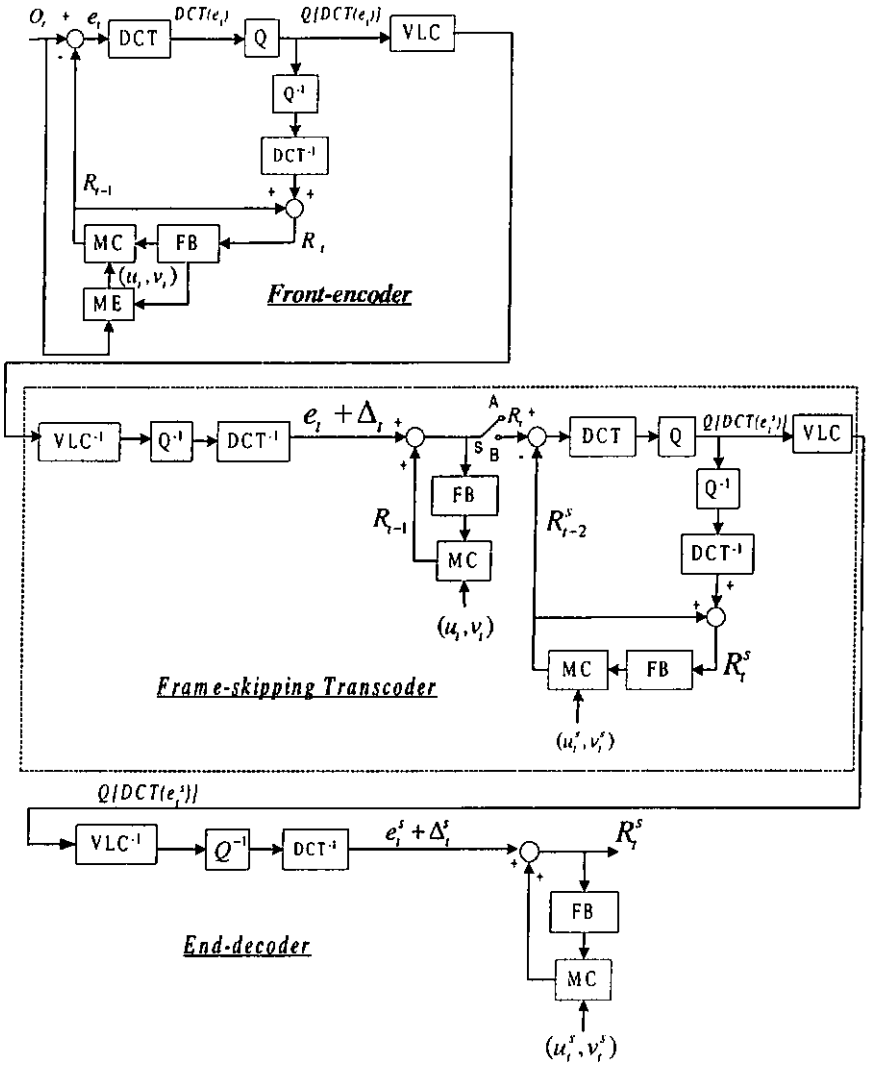


Figure 3.2 Frame-skipping transcoder in pixel-domain.

Table 3.1. Switch position for different modes of frame skipping.

Frame skipping mode	*S* Position
Skipped frame	*A*
Non-skipped frame	*B*

Assume that frame *t-1*, R_{t-1}, is skipped. However, R_{t-1} is required to act as the reference frame for the reconstruction of frame *t*, R_t, such that

$$R_t(i,j) = R_{t-1}(i+u_t, j+v_t) + e_t(i,j) + \Delta_t(i,j) \tag{3.3}$$

where $\Delta_t(i,j)$ represents the reconstruction error of the current frame in the front-encoder due to the quantization, and $e_t(i,j)$ is the residual signal between the current frame and the motion-compensated frame,

$$e_t(i,j) = O_t(i,j) - R_{t-1}(i+u_t, j+v_t) \tag{3.4}$$

Substituting (4) into (3), we obtain the expression for R_t,

$$R_t(i,j) = O_t(i,j) + \Delta_t(i,j) \tag{3.5}$$

In the transcoder, an optimized motion vector for the outgoing bitstream can be obtained by applying the motion estimation such that

$$(u_t^s, v_t^s) = \arg \min_{(m,n)\in S} SAD^s(m,n) \tag{3.6}$$

$$SAD^s(m,n) = \sum_i^M \sum_j^N |R_t(i,j) - R_{t-2}^s(i+m, j+n)| \tag{3.7}$$

where $R_{t-2}^s(i,j)$ denotes a reconstructed pixel in the previous non-skipped reference frame. The superscript "s" is used to denote the symbol after performing the frame-skipping transcoder. Although the optimized motion vector can be obtained by a new motion estimation, it is not desirable because of its high computational complexity. Reuse of the incoming motion vectors has been widely accepted because it is considered to be almost as good as performing a new full-scale motion estimation and was assumed in many transcoder architectures[1,12]. Thus, we assume that the new motion vector is (u_t^s, v_t^s). Hence, the reconstructed pixel in the current frame after the end-decoder is,

$$R_t^s(i,j) = R_{t-2}^s(i+u_t^s, j+v_t^s) + e_t^s(i,j) + \Delta_t^s(i,j) \tag{3.8}$$

where $e_t^s(i,j) = R_t(i,j) - R_{t-2}^s(i+u_t^s, j+v_t^s)$ and $\Delta_t^s(i,j)$ represents the requantization error due to the re-encoding in the transcoder, then,

$$R_t^s(i,j) = O_t(i,j) + \Delta_t(i,j) + \Delta_t^s(i,j) \tag{3.9}$$

This equation implies that the reconstructed quality of the non-skipped frame deviates from the input sequence to the transcoder, R_t. An additional error, Δ_t^s, is introduced. Re-encoding of the current frame involves a re-computation of the residual signal between the current frame and the non-skipped reference frame. Note that frame *t-2* acts as the reference instead of frame *t-1*, since frame *t-1* does not exist after frame skipping. The newly quantized DCT-domain data are then re-computed by means of the DCT and quantization processes. This re-encoding procedure can lead to an additional error Δ_t^r. The effect of the re-encoding error is depicted in Figure 3.3 where the "Salesman" sequence was transcoded at half of the incoming frame-rate. In the figure, the peak signal-to-noise ratio (PSNR) of the frame-skipping pictures is plotted to compare with that of the same pictures directly using a decoder without a transcoder. This figure shows that the re-encoding error leads to a drop in picture quality of about 3.5dB on average, which is a significant degradation. Details on the simulation environment and coding parameters used in the simulation are described in Section 3.5.

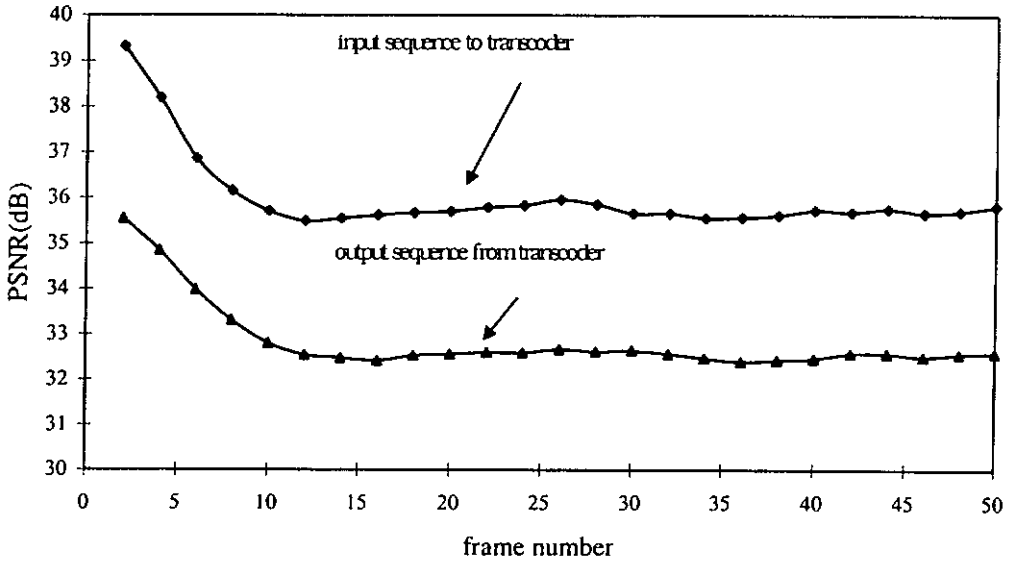


Figure 3.3. Quality degradation of conventional frame-skipping transcoder for the "Salesman" sequence.

3.3 Low-complexity Frame-skipping for High Performance Video Transcoding

Besides the quality issue mentioned above, the pixel-domain transcoder also has a high processing complexity. This is due to the fact that the skipped frame must be decompressed completely, and should act as the reference frame to the non-skipped frame for reconstruction. In this section, we present a new architecture for frame-rate reduction in order to achieve an improved picture quality and a reduced complexity of the transcoded sequence.

The architecture of the proposed transcoder is shown in Figure 3.4. The input bitstream is first parsed with a variable-length decoder to extract the header information, coding mode, motion vectors and quantized DCT coefficients for each macroblock. Each macroblock is then manipulated independently. The two switches *S1* and *S2* are employed to update the DCT-domain buffer for the transformed and quantized residual signal depending on the coding mode originally used at the front encoder for the current

macroblock being processed. The switch positions for different coding modes are shown in Table 3.2. When the macroblock is not motion compensated, the previous residual signal in the DCT-domain is directly fed back from the DCT-domain buffer to the summer, and the sum of the input residual signal and the previous residual signal in the DCT-domain is updated in the buffer. Note that all operations are performed in the DCT-domain, thus the complexity of the frame-skipping transcoder is reduced. Also, the quality degradation of the transcoder introduced by Δ_t^s is avoided. When motion compensation is used, motion compensation, DCT, inverse DCT, quantization and inverse quantization modules are activated to update the DCT-domain buffer. The advantages of this DCT-domain buffer arrangement and the details of our method are described in the following subsections. Note that the switch *S3* is used to control the frame rate and refresh the frame buffer for non-skipped frames. Table3.3 shows the frame-skipping modes of our proposed transcoder.

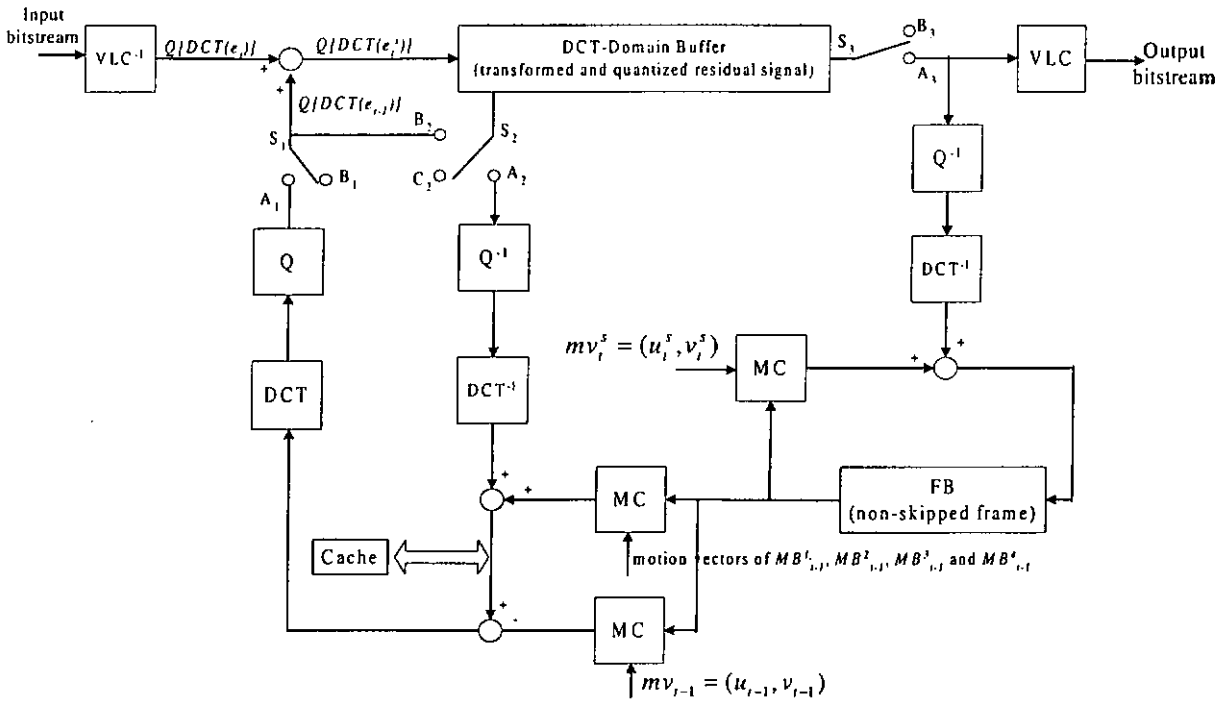


Figure 3.4. The proposed frame-skipping transcoder.

Table 3.2. Different coding modes for switches S1 and S2.

Coding mode	*S1* Position	*S2* Position
No MC	B_1	B_2
MC	A_1	A_2

Table 3.3 Switch positions for different frame-skipping modes of our proposed transcoder.

Frame-skipping mode	*S3* Position
Skipped frame	B_3
Non-skipped frame	A_3

3.3.1 Direct summation of DCT coefficients for macroblock without motion compensation

For those macroblocks coded without motion compensation, the direct summation of DCT coefficients is employed such that the DCT transform pair and motion compensation operation are not needed. For typical video conferencing, a majority of the video signal is coded without motion compensation, and hence the complexity reduction realised by using the direct summation is significant. In figure 3.5, a situation in which

one frame is dropped is illustrated. We assume that MB_t represents the current macroblock and MB_{t-1} represents the best matching macroblock to MB_t. Since MB_t is coded without motion compensation, the spatial position of MB_{t-1} is the same as that of MB_t, and MB_{t-2} represents the best matching macroblock to MB_{t-1}. Since R_{t-1} is dropped, for MB_t, we need to compute a motion vector, (u_t^s, v_t^s), and the prediction error in DCT-domain, $Q[DCT(e_t^s)]$, by using R_{t-2} as a reference. Since the motion vector in MB_t is zero, then

$$(u_t^s, v_t^s) = (u_{t-1}, v_{t-1}) \tag{3.10}$$

Since re-encoding can lead to an additional error, it could be avoided if $Q[DCT(e_t^s)]$ can be computed in the DCT-domain.

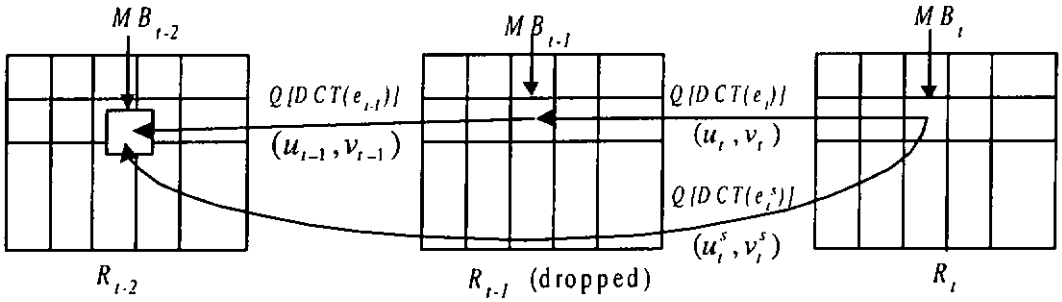


Figure 3.5 Residual signal re-computation of frame skipping for macroblocks without motion compensation.

In figure 3.5, the pixels in MB_t can be reconstructed by performing inverse quantization and inverse DCT of $Q[DCT(e_t)]$ and summing this residual signal to pixels in MB_{t-1} which can be similarly reconstructed by performing inverse quantization and inverse DCT of $Q[DCT(e_{t-1})]$ and summing this residual signal to pixels in the corresponding MB_{t-2}. The reconstructed macroblocks MB_t and MB_{t-1} are given by,

$$MB_t = MB_{t-1} + e_t + \Delta_t \tag{3.11}$$

and

$$MB_{t-1} = MB_{t-2} + e_{t-1} + \Delta_{t-1} \tag{3.12}$$

Note that Δ_t and Δ_{t-1} represents the quantization error of MB_t and MB_{t-1} in the front-encoder. Substituting (3.12) into (3.11), (3.13) is obtained

$$e_t^s = (e_t + \Delta_t) + (e_{t-1} + \Delta_{t-1}) \tag{3.13}$$

where $e_t^s = MB_t - MB_{t-2}$. This is the prediction error between the current macroblock and its corresponding reference macroblock. By applying the DCT for e_t^s and taking into account the linearity of DCT, we obtain the expression of e_t^s in the DCT-domain.

$$DCT(e_t^s) = DCT(e_t + \Delta_t) + DCT(e_{t-1} + \Delta_{t-1}) \tag{3.14}$$

Then the newly quantized DCT coefficients of prediction error are given by

$$Q[DCT(e_t^s)] = Q[DCT(e_t + \Delta_t) + DCT(e_{t-1} + \Delta_{t-1})] \tag{3.15}$$

Note that, in general, quantization is not a linear operation because of the integer truncation. However, $DCT(e_t + \Delta_t)$ and $DCT(e_{t-1} + \Delta_{t-1})$ are the output from the quantizer and they are divisible by the quantizer step-size. Thus,

$$Q[DCT(e_t^s)] = Q[DCT(e_t + \Delta_t)] + Q[DCT(e_{t-1} + \Delta_{t-1})] \tag{3.16}$$

Since Δ_t and Δ_{t-1} are introduced due to the quantization at the front encoder, $Q[DCT(e_t + \Delta_t)] = Q[DCT(e_t)]$ and $Q[DCT(e_{t-1} + \Delta_{t-1})] = Q[DCT(e_{t-1})]$. We obtain the final expression of the prediction error in the quantized DCT-domain by using R_{t-2} as a reference,

$$Q[DCT(e_t^s)] = Q[DCT(e_t)] + Q[DCT(e_{t-1})] \tag{3.17}$$

Equation (3.17) implies that the newly quantized DCT coefficient $Q[DCT(e_t^s)]$ can be computed in the DCT-domain by summing directly the quantized DCT coefficients

between the data in the DCT-domain buffer and the incoming DCT coefficients, whilst the updated DCT coefficients are stored in the DCT-domain buffer, as depicted in figure 3.4, when switches S_1 and S_2 are connected to B_1 and B_2 respectively. Since it is not necessary to perform motion compensation, DCT, quantization, inverse DCT and inverse quantization, the complexity is reduced. Furthermore, since requantization is not necessary for this type of macroblock, the quality degradation of the transcoder introduced by Δ_t^s is also avoided. Figure 3.6 shows the distribution of the coding mode for the typical "salesman" sequence; it is clear that over 95% of the macroblocks are coded without motion compensation. By using a direct summation of DCT coefficients for non-moving macroblocks, the computational complexity involved in processing these macroblocks can be reduced significantly and the additional re-encoding error can be avoided.

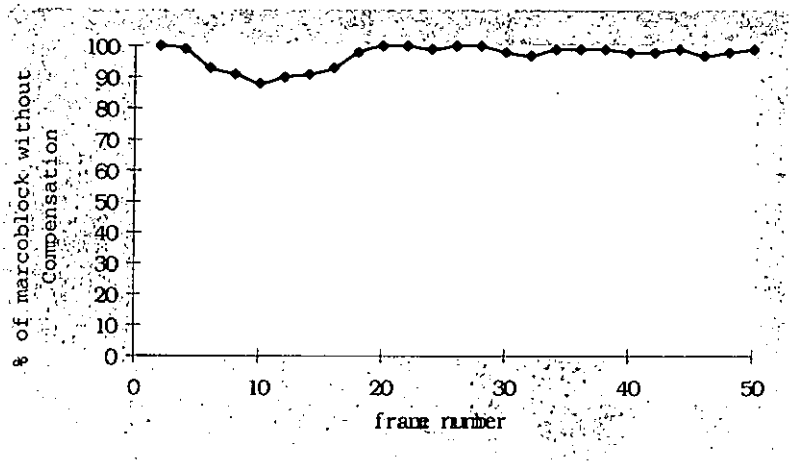


Figure 3.6 Distribution of coding modes for "salesman" sequence.

3.3.2 DCT-domain buffer updating for motion-compensated macroblock

The use of direct summation of DCT coefficients for macroblocks without motion compensation reduces the complexity and eliminates the re-encoding error. This

advantage stems from the use of the DCT-domain buffer. For motion-compensated macroblocks, direct summation cannot be employed since MB_{t-1} is not on a macroblock boundary, as depicted in figure 3.7. In other words, $Q[DCT(e_{t-1})]$ is not available from the incoming bitstream. It is possible to use the motion vectors and quantized DCT coefficients of the four neighboring macroblocks with MB_{t-1}, MB^1_{t-1}, MB^2_{t-1}, MB^3_{t-1} and MB^4_{t-1}, to come up with $Q[DCT(e_{t-1})]$.

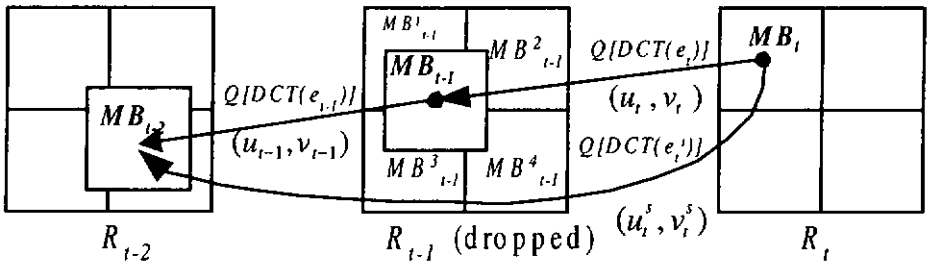


Figure 3.7 Residual signal re-computation of frame-skipping for motion-compensated macroblocks.

First, inverse quantization and inverse DCT of the quantized DCT coefficients of MB^1_{t-1}, MB^2_{t-1}, MB^3_{t-1} and MB^4_{t-1} are performed to obtain their corresponding prediction errors in the pixel-domain. The MB_{t-1} is composed of four components as shown in figure 3.8. Thus, each segment of the reconstructed pixels in MB_{t-1} can be obtained by summing its prediction errors and its motion-compensated segment of the previous non-skipped frame stored in the frame buffer, as shown in the block diagram of figure 3.4.

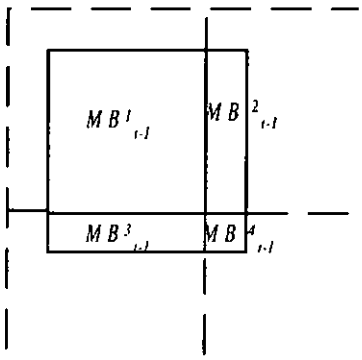


Figure 3.8 Composition of MB_{t-1}.

After all pixels in MB_{t-1} have been reconstructed, we need to find the prediction error, e_{t-1}. Actually, e_{t-1} is equal to the reconstructed pixel in MB_{t-1} subtracted from the motion-compensated macroblock from the previous non-skipped frame stored in the frame buffer. In order to obtain the motion-compensated macroblock, we need to find a motion vector of MB_{t-1}. Again, MB_{t-1} is not on a macroblock boundary; it is possible to use the bilinear interpolation from the motion vectors mv_{t-1,MB_1}, mv_{t-1,MB_2}, mv_{t-1,MB_3} and mv_{t-1,MB_4} of the four neighboring macroblocks with MB_{t-1}, MB^1_{t-1}, MB^2_{t-1}, MB^3_{t-1} and MB^4_{t-1}, to come up with an approximation of mv_{t-1} [13]. However, the bilinear interpolation of motion vectors has several drawbacks [1,12]. Thus, the dominant vector selection approach is used [1,12] to select one dominant motion vector from the four neighboring macroblocks. A dominant motion vector is defined as the motion vector carried by a dominant macroblock. The dominant macroblock is the macroblock that has the largest overlapped segment with MB_{t-1}.

Hence, e_{t-1} can be computed and it is transformed and quantized to $Q[DCT(e_{t-1})]$. Since quantization is performed in the formation of $Q[DCT(e_{t-1})]$, some quantization

error, Δ_{t-1}^{s}, is introduced. The newly quantized DCT coefficient $Q[DCT(e_t^s)]$ of a motion-compensated macroblock can be computed by

$$Q[DCT(e_t^s)] = Q[DCT(e_t)] + Q[DCT(e_{t-1})] + Q[DCT(\Delta_{t-1}^{s})] \qquad (3.18)$$

In most cases, Δ_{t-1}^{s} is smaller than Δ_t^s since the distance to the reference frame is one frame less. The evidence will be shown in Section 3.5.

In order to reduce the implementation complexity of the motion-compensated macroblock, a cache subsystem is added to our proposed transcoder, as depicted in figure 3.4. Since motion compensation of multiple macroblocks may require the same pixel data, a cache subsystem is implemented to reduce redundant inverse quantization, inverse DCT and motion compensation computations. We have found that the arrangement is significant since the frequency of caching hits is high. This is due to the fact that the locality of motion is often present within each frame.

3.3.3 Multiple Frame-skipping in our Proposed Transcoder

Another advantage of the proposed frame-skipping transcoder is that when multiple frames are dropped, it can be processed in the forward order, thus eliminating the multiple DCT-domain buffers that are needed to store the incoming quantized DCT coefficients of all dropped frames. Figure 3.9 shows a scenario in which two frames are dropped. When R_{t-2} is dropped, we store the DCT coefficients of its prediction errors in the DCT-domain buffer. The stored DCT coefficients of prediction errors will be used to update the DCT coefficients of prediction errors at the next dropped frame. This means that when R_{t-1} is dropped, our proposed scheme updates the DCT coefficients of prediction errors for each macroblock according to its coding mode.

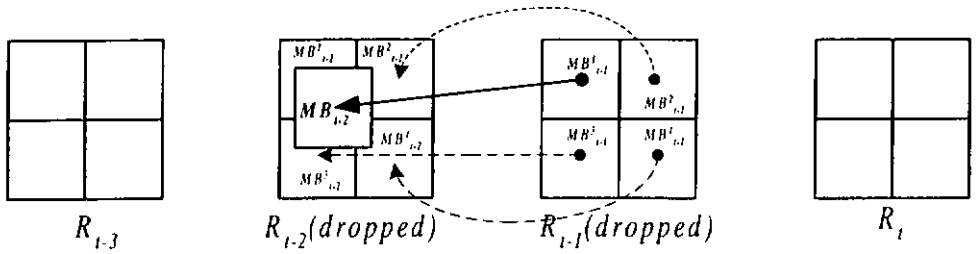


Figure 3.9 Multiple frame skipping of our proposed transcoder.

For example, macroblocks, MB^2_{t-1}, MB^3_{t-1}, and MB^4_{t-1}, in R_{t-1} are coded without motion compensation. From eqn (3.17), the DCT coefficients of prediction errors in the DCT-domain buffer are added to the corresponding incoming prediction errors of the macroblock in R_{t-1}. The buffer is then updated with the new DCT coefficients. In figure 3.9, MB^1_{t-1} is a motion-compensated macroblock. It is necessary to perform the re-encoding of the macroblock pointed by MB^1_{t-1} and then add the corresponding incoming DCT-coefficients to form the updated data in the DCT-domain buffer, as indicated in eqn(3.18). By using our proposed scheme, only one DCT-domain buffer is needed for all the dropped frames. The flexibility of multiple frame-skipping provides the fundamental framework for dynamic frame-skipping, which is used in multipoint video conferencing.

3.4 Dynamic Frame Allocation for Video Combining in Multipoint Conferencing

In a multipoint video conferencing system, usually only one or two participants are active at any given time [7]. The active conferees need higher frame rates to produce a better video quality as well as to present a smoother motion. Figure 3.10 shows the proposed system architecture for video combiner in multipoint video conferencing. Our approach for video combining is based on frame-skipping transcoding which primarily consists of the DCT-domain approach without the requirement of the asymmetric

network channels. Thus, the original video sequence can be encoded by fully utilizing the available channel bandwidth. Up to four QCIF video bitstreams are received by the video combiner from the conference participants. Each QCIF bitstream is processed by our proposed frame-skipping transcoder. The main function of a frame-skipping transcoder is frame-rate reduction. Note that the output frame rates from the four transcoders are not constant, and they are not necessarily equal to one another.

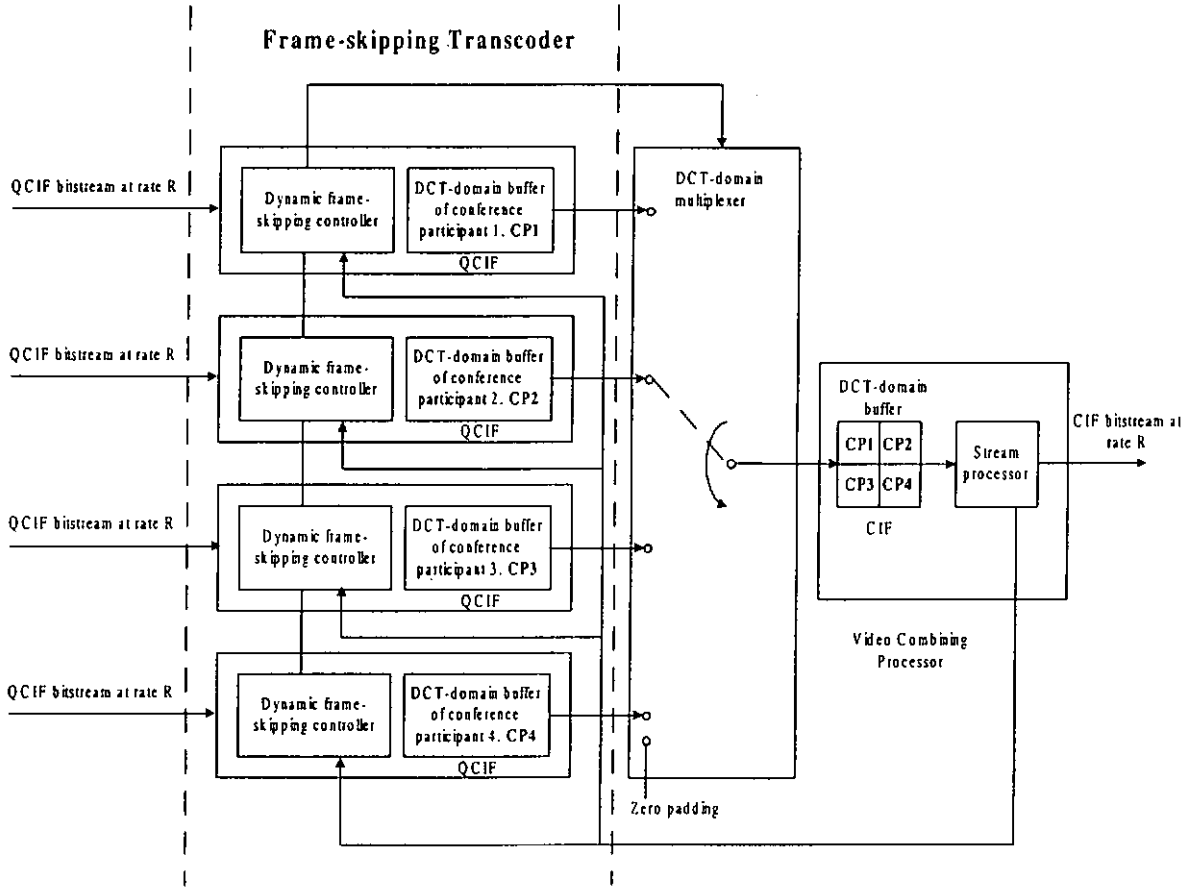


Figure 3.10 System architecture for video combiner using the frame-skipping transcoder.

In the transcoder, the frame-skipping controller can dynamically distribute the encoded frames to each sub-sequence by considering their motion activities so that the quality of the transcoded video can be improved. The frame rate required to transcode a sub-sequence is highly related to its motion activity [13,51-52]. Thus, it is necessary to

regulate the frame rate of the each transcoder according to the motion activity in the current frame (MA_t) of the sub-sequence. To obtain a quantitative measure for MA_t, we use the accumulated magnitudes of all of the motion vectors estimated for the macroblocks in the current frame [8,24], i.e.,

$$MA_t = \sum_{i=1}^{N} \left|(u_t^s)_i\right| + \left|(v_t^s)_i\right| \tag{3.19}$$

where N is the total number of macroblocks in the current frame, and $(u_t^s)_i$ and $(v_t^s)_i$ are the horizontal and vertical components of the motion vector of the ith macroblock, which uses the previous non-skipped frame as a reference.

If the value of MA_t after a non-skipped frame exceeds the predefined threshold, T_{MA}, the incoming frame should be kept. It is interesting to note that the T_{MA} is set according to the outgoing bit rate of the video combiner, but this is not the focus of this paper. By adaptively adjusting the frame rate of each sub-sequence according to the MA_t, the proposed architecture can allocate more frames for a sub-sequence with high motion activity and less frames for a sub-sequence with low motion activity.

Our proposed transcoder updates the quantized DCT coefficients of the current frame in the DCT-domain buffer for each QCIF sub-sequence. At the beginning of the formation of a combining sequence, a decision is made as to which DCT-domain buffers should be included in the new buffer by the dynamic frame-skipping controllers. If the current frame of the sub-sequence is selected, it is picked out by the multiplexer and the DCT coefficients in the corresponding DCT-domain buffer are copied in the video combining processor's buffer with the size of CIF. The quantized DCT coefficients of each conference participant only need to be mapped according to figure 3.10. Hence, the DCT-domain buffer of the conference participant 1, *CP1*, is mapped to the first quadrant

of the combined picture, the DCT-domain buffer of the conference participant 2, *CP2*, is mapped to the second quadrant, etc. If the current frame of the sub-sequence is skipped, the corresponding quadrant is filled with zero value. Following the assembly of the new DCT buffer in the video combining processor, the data in the buffer are coded in compressed bitstream by the stream processor.

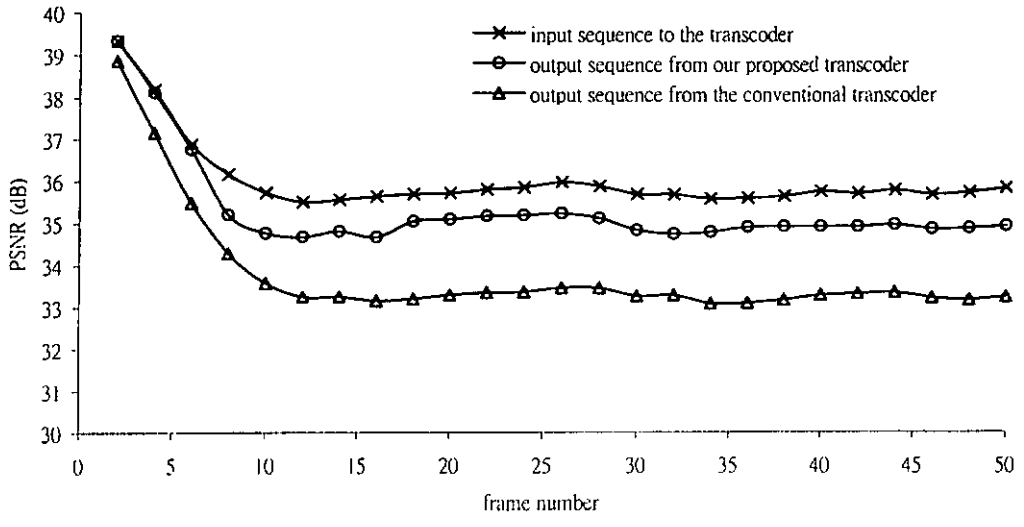
3.5 Simulation Results

In this section, we present some simulation results. A series of computer simulations were conducted to evaluate the overall efficiency of the proposed frame-skipping transcoder. The performance of the proposed video combiner for multipoint continuous presence video conferencing is also presented below.

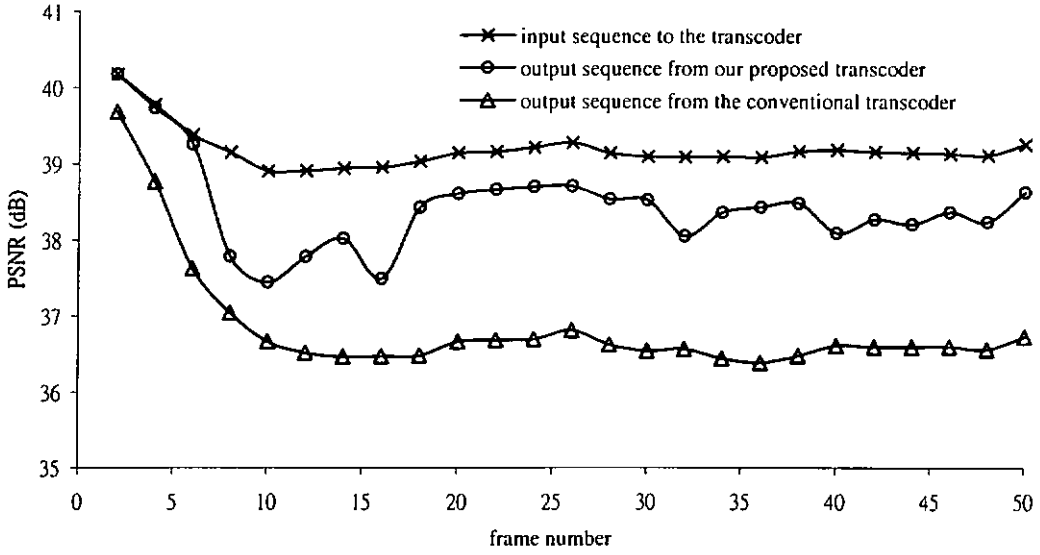
3.5.1 Performance of the Frame-Skipping Transcoder

To evaluate the overall efficiency of the proposed frame-skipping transcoding approach, all test sequences of QCIF (176×144) were encoded at high bitrate (64kb/s and 128kb/s) using a fixed quantization parameter. At the front encoder, the first frame was coded as intraframe (I-frame), and the remaining frames were encoded as interframes (P-frames). These picture-coding modes were preserved during the transcoding. The PSNR performance of the proposed frame-skipping transcoder for the "Salesman" sequence is shown in figure 3.11. At the front encoder, the original test sequence "Salesman" was encoded at 64kb/s and 128kb/s in figure 3.11(a) and figure 3.11(b), respectively, and then transcoded into 32kb/s and 64kb/s at half of the incoming frame rate. As shown in figure 3.11, the proposed transcoder outperforms the conventional pixel-domain transcoder. Also, table 3.4 shows that it has a speed-up of about 7 times faster than that of the conventional transcoder for the "Salesman" sequence. This is

because the probability of the macroblock coded without motion compensation happens more frequently in typical sequences, and this type of macroblock should not introduce any re-encoding error due to the direct summation of the DCT coefficients. As shown in table 3.5, the average PSNR performance of the macroblock coded without motion compensation in our proposed transcoder is significantly better than that of the conventional transcoder. Thus, we can achieve significant computational savings while maintaining a good video quality on these macroblocks. On the other hand, our proposed transcoder also shows an improvement with respect to the motion-compensated macroblock, as depicted in table 3.5. This is due to the fact that Δ_{t-1}^{s} is smaller than Δ_{t}^{s} in most cases, as mentioned in Equation (3.18). Furthermore, the cache system in the transcoder can reduce the computational burden of re-encoding the motion-compensated macroblocks. All these advantages combined gives rise to significant computational saving as well as quality improvement. These demonstrate the effectiveness of the proposed frame-skipping transcoder. The simulation results of other test sequences are summarized in table 3.4 and table 3.5



(a)



(b)

Figure 3.11 Performance of the proposed transcoder of "Salesman" sequence encoded at (a) 64kb/s with 30 frames/s, and then transcoded to 32kb/s with 15 frames/s. (b) 128Kb/s with 30 frames/s, which are then transcoded to 64kb/s with 15 frames/s.

Table3.4. Speed-up ratio of the proposed transcoder. The frame-rate of incoming bitstream is 30 frames/s which are then transcoded to 15 frames/s.

Sequences	Input bitrate	Speed-up ratio
Salesman	64k	6.75
	128k	7.64
News	64k	6.08
	128k	6.68
Hall	64k	3.57
	128k	3.96

Table 3.5. Performance of the proposed transcoder. The frame-rate of incoming bitstream is 30 frames/s which are then transcoded to 15 frames/s.

Sequences	Input bitrate	Conventional transcoder			Our proposed transcoder		
		MC region	Non-MC region	All region	MC region	Non-MC region	All region
Salesman	64k	30.25	34.31	33.77	31.45	36.32	35.47
	128k	33.58	37.14	36.85	34.11	39.28	38.62
News	64k	30.44	34.66	34.03	31.37	36.49	35.55
	128k	34.07	37.47	37.13	34.56	39.79	38.97
Hall	64k	36.14	37.07	36.93	36.86	37.79	37.06
	128k	38.3	39.43	38.94	38.36	40.23	39.27

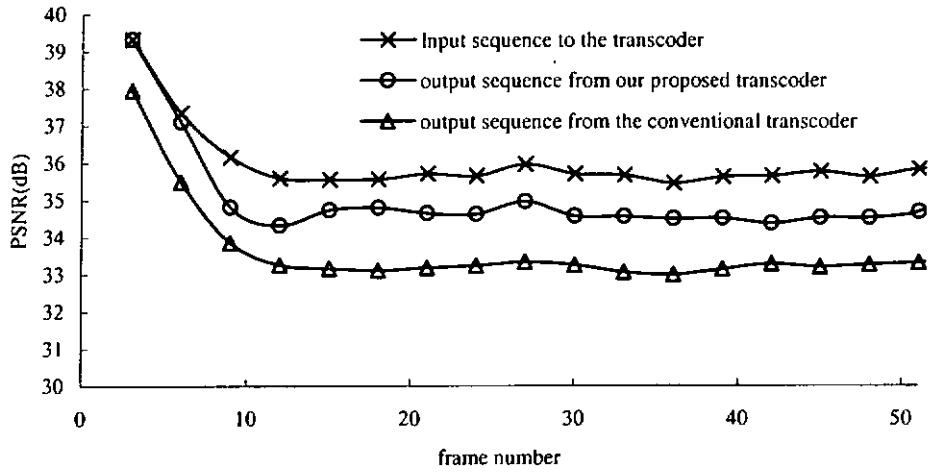
In order to illustrate the effects of the proposed frame-skipping transcoder with multiple-frame dropping, Table 3.6, Table 3.7 and Figure 3.12 set forth the results of the frame-skipping transcoding for which the frames are temporally dropped by a factor of *2*. The results appear to be similar to that of the above. But it is quite apparent that the conventional frame-skipping transcoder gives the worst performance, and our proposed transcoder provides a significant improvement. Also the computational complexity is reduced remarkably.

Table 3.6 Speed-up ratio of the proposed transcoder. The frame-rate of incoming bitstream is 30 frames/s which are then transcoded to 10 frames/s.

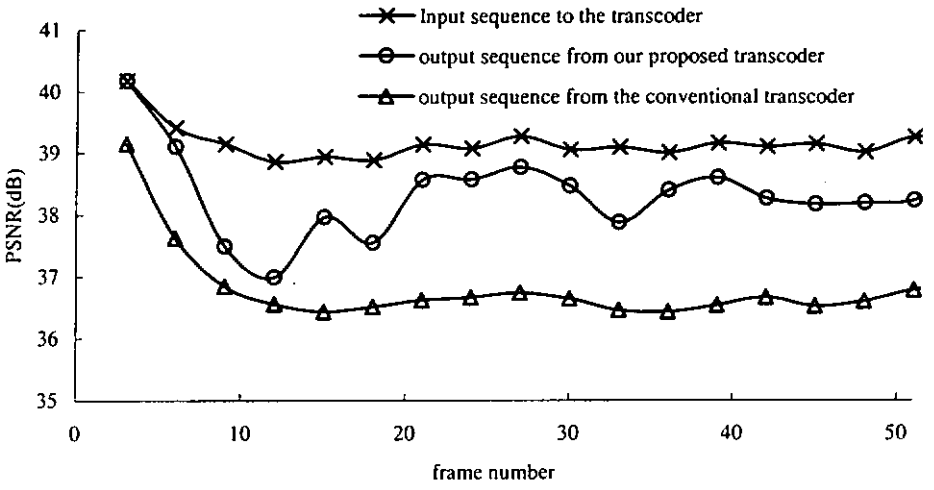
Sequences	Input bitrate	Speed-up ratio
Salesman	64k	9.78
	128k	11.82
News	64k	7.68
	128k	8.84
Hall	64k	4.38
	128k	4.69

Table 3.7 Performance of the proposed transcoder. The frame-rate of incoming bitstream is 30 frames/s which are then transcoded to 10 frames/s.

Sequences	Input bitrate	Conventional transcoder			Our proposed transcoder		
		MC region	Non-MC region	All region	MC region	Non-MC region	All region
Salesman	64k	29.85	34.12	33.45	30.89	35.89	35.30
	128k	33.59	37.03	36.70	33.68	38.95	38.44
News	64k	30.04	34.49	33.69	30.82	36.23	35.3
	128k	34.11	37.34	36.93	35.23	39.36	38.62
Hall	64k	34.2	36.37	35.96	36.74	37.75	37.01
	128k	38.2	38.81	38.52	38.21	39.72	38.91



(a)



(b)

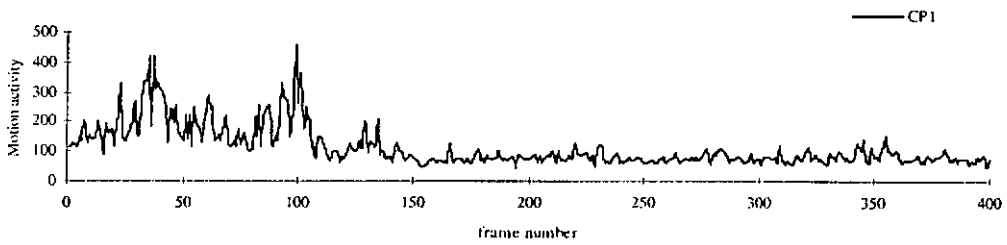
Figure 3.12. Performance of the proposed transcoder of "Salesman" sequence encoded at (a) 64Kb/s with 30 frames/s, and then transcoded to 21kb/s with 10 frames/s. (b) 128kb/s with 30 frames/s, which are then transcoded to 42kb/s with 10 frames/s.

3.5.2 Performance of Continuous Presence Video Conferencing System

For our simulation, we recorded a four-point video conferencing session. Each conferee's video was encoded into a QCIF format at 128kb/s, as shown in Figure 3.13. The four video sequences were transcoded and combined into a CIF format. We then selected segments of the combined sequence to form a 400-frame video sequence in which the first person was most active in the first 100 frames, the second person was

most active in the second 100 frames, and so on. The motion activities for the four conference participants and the combined video sequence are shown in Figure 3.14. In the figure, the top four curves correspond to the four participants in the upper left, upper right, lower left and lower right corners, respectively. The bottom curve represents the motion activity of the combined video sequence. The figure shows that although there were short periods of time when multiple participants were active, only one participant was active during most of the time, while other participants were relatively inactive. The overall motion activity of the combined video sequence is relatively random. This indicates that multipoint video conferencing is a suitable environment for dynamic allocation of the encoding frames to each participant.

Figure 3.13 Encoded frame 194 of the four conferee's videos, which are received by the MCU.



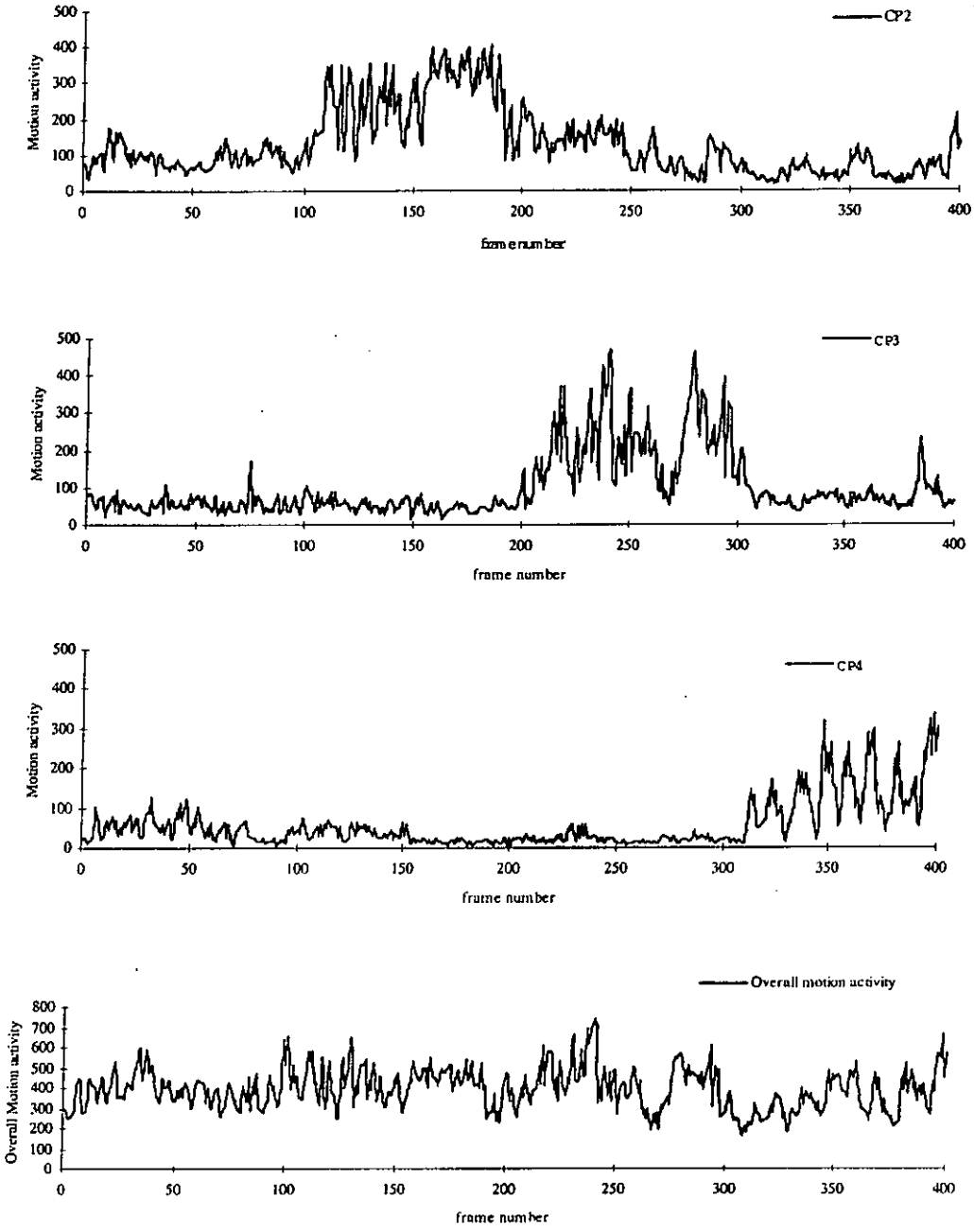
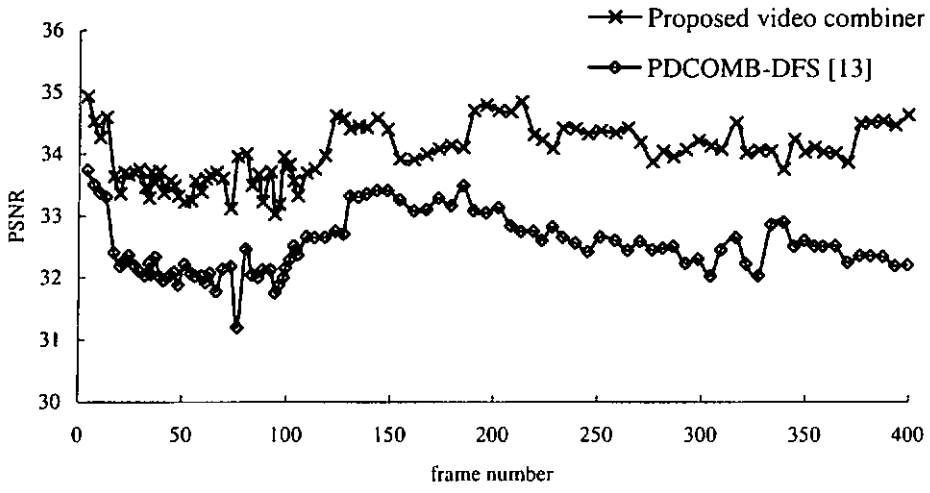


Figure 3.14 Motion activity of a multipoint videoconference.

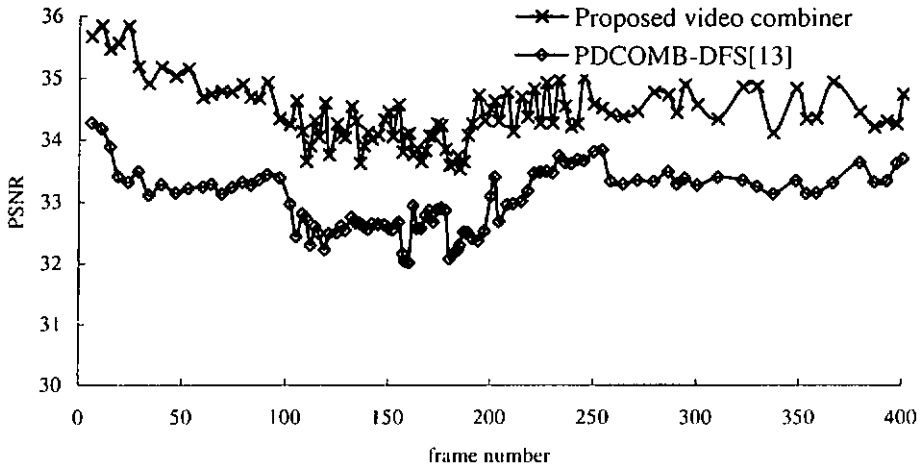
In the following discussion, we will analyse the performance of the proposed video combiner for continuous presence multipoint video conferencing system as compared to the conventional pixel-domain combiner with dynamic frame-skipping (PDCOMB-DFS) [13]. Since the active participants need higher frame-rates to produce

an acceptable video quality while the inactive participants only need lower frame-rates to produce an acceptable quality, we use dynamic frame allocation in both the proposed video combiner and the PDCOMB-DFS to distribute the encoded frames into each sub-sequence according to the motion activities. Inevitably, this improvement is made by sacrificing a certain amount of quality of the motion inactive periods. Figure 3.15 shows the PSNR performances of the conference participants for different video combining approaches. For example, the participant is most active from frame 0 to frame 100 in Figure 3.15(a) (as shown in the motion activity plot in Figure 3.14). This active period is transcoded more frequently following the motion activities of the sub-sequence, therefore the videos displayed on the receiver are smoother. It can be seen from Figure 3.15 that the conventional PDCOMB-DFS loses due to the double-encoding aspect and the proposed video combiner offers a much better quality as compared to the PDCOMB-DFS. The gain can be as high as 1.5-2.0 dB for both the active and non-active periods due to the high efficiency of our proposed frame-skipping transcoder. A frame (194^{th} frame) in the combined sequence is shown in Figure 3.16. It can be seen that the video quality of the active participant (at the upper right corner) with the proposed video combiner is much better than that with the PDCOMB-DFS. Due to the dynamic frame allocation based on the motion activities, the degradation of video quality of the inactive participants is not very visible and this is also supported by Figure 3.13 and Figure 3.16. The PSNR's of each participant in the overall video sequence at 128 kb/s using different video combiners are summarized in Table 3.8. The diagonal PSNR's indicate more active motion in different time slots of individual conference participants. The table shows that by using the proposed video combiner the PSNR's among all conference

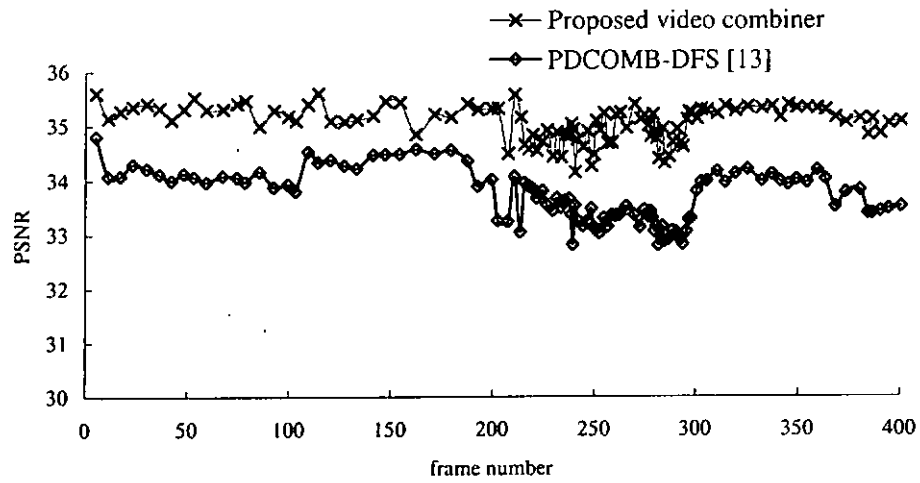
participants are much improved as compared to the PDCOMB-DFS during both the active and non-active periods. In practical multipoint video conferencing, active participants are given most attention. Improvement of the video quality of the active participants is particularly important and we have shown that the proposed video combiner can achieve a significant improvement.



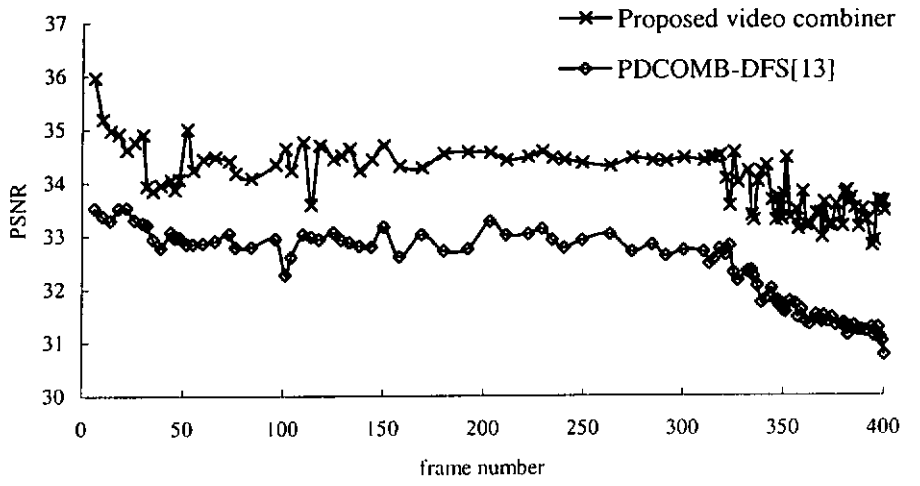
(a)



(b)



(c)



(d)

Figure 3.15 PSNR performance of a conference participant who is most active (a) between frame 0 and frame 100, (b) between frame 101 and frame 200, (c) between frame 201 and frame 300, (d) between frame 301 and frame 400.

Figure 3.16 Frame 194 of the combined video sequence using (a) PDCOMB-DFS [13] (b) our video combiner using the proposed frame-skipping transcoder. The active conference participant is at the upper right corner.

Table 3.8. Average PSNR's of the combined video sequence.

	Frame 1 – 100		Frame 101 - 200		Frame 201 - 300		Frame 301 - 400	
	A	B	A	B	A	B	A	B
1st conference participant (most active during frame 1 –100)	**32.21**	**33.65**	32.99	34.15	32.60	34.30	32.41	34.19
2nd conference participant (most active during frame 101 –200)	33.42	35.07	**32.55**	**34.07**	33.39	34.55	33.36	34.50
3rd conference participant (most active during frame 201 –300)	34.11	35.31	34.31	35.24	**33.32**	**34.84**	33.85	35.20
4th conference participant (most active during frame 301 –400)	33.08	34.48	32.84	34.43	32.91	34.44	**31.66**	**33.64**

A - PDCOMB-DFS [13].
B - Proposed video combiner.

3.6 Conclusions

This chapter proposes a low-complexity and high quality frame-skipping transcoder. Its low complexity is achieved by: 1) a direct summation of the DCT coefficients for macroblocks coded without motion compensation to deactivate most complex modules of the transcoder, and 2) a cache subsystem for motion-compensated macroblocks to reduce redundant IDCT and inverse quantization. We have also shown that a direct summation of the DCT coefficients can eliminate the re-encoding error due to requantization. Furthermore, our proposed frame-skipping transcoder can be processed in the forward order when multiple frames are dropped. Thus, only one DCT-domain buffer is needed to store the updated DCT coefficients of all dropped frames. Overall, the proposed frame-skipping transcoder produces a better picture quality than the conventional frame-skipping transcoder at the same reduced bitrates.

We have also integrated our proposed frame-skipping transcoder into a new video combining architecture for continuous presence multipoint video conferencing. In multipoint video conferencing, usually only one or two participants are active at any given time. When the frame skipping transcoding approach is used, the frame rate of coding a sub-sequence needed to achieve a certain quality level depends very much on its

motion activity, using the frame-skipping transcoding approach. We can achieve a better video quality by dynamic frame allocation based on the motion activities of the sub-sequences. Since re-encoding is minimized in our frame-skipping transcoder, the proposed architecture provides a better performance than a conventional video combiner in terms of quality and complexity. However, many problems still remain to be investigated. For example, it would be desirable to design an efficient global frame allocation algorithm which can guarantee a reasonable sub-optimality frame-rate for all sub-sequences in order to fulfill the desired output bitrate of the video combiner. Nevertheless, it can be seen that a video combiner using the frame-skipping transcoding approach is able to provide a new and viable continuous presence multipoint video conferencing service in the near future.

Chapter 4

Region-based Object Tracking for Multipoint Video Conferencing using Wavelet Transform

In this chapter, a wavelet-based video coder include: 1) a user-specified region of interest selection as to which the region can be changed by the user at any I-frame; 2) a dynamic region tracking technique by which the video is tracked and updated according to motion activity and 3) an adaptive bit allocation that allows the user to specify the relative quality between the foreground and the background will be described. This architecture guarantees a high video quality in the region of interest while reducing the overall bit rate and the computation time. Experimental results confirm that the approach produces a good video quality even under low bit rates.

4.1. Introduction

With the advance of video compression and networking technologies, multipoint video conferencing becomes popular in the consumer market [15]. Most video conferencing systems use DCT-based encoders. A good performance can be achieved with a large bandwidth [53]. However, under low bit rates, the DCT-based encoder exhibits visually annoying blocking artifacts. Recently, wavelets have been used in internet applications. The major advantage of using a wavelet is its high quality and the absence of blocking artifacts when compared to the conventional video encoder [54-55]. Although a wavelet-based coder can achieve good quality, its computational speed is an area of concern. A way to speed up the computation is to explore the fact that the various regions in an image are not of equal importance. This concept has been adopted in dynamic bit allocation and frame-skipping technique [56].

In this chapter, a new region-based video coder architecture is proposed to achieve a good video quality with a low complexity. The proposed video coder is based on the adaptive region-based updating technique by which the video is updated according to the motion activity. This architecture allows a high quality video in the region of interest while reducing the overall bit rate and computation time. Since the user might be in fast motion when active, a simple and fast object tracking technique is proposed to locate the region of interest. This approach produces a good video quality even under low bit rates.

4.2. The Proposed architecture

Figure 4.1 shows the system architecture for the proposed video coder in multipoint video conferencing. Since the video encoder is wavelet-based, blocking artifacts are avoided. Our proposed region-based video encoder has two major features:

1) selection of the region of interest; and

2) adaptive bit allocation for the foreground (region of interest) and the background.

The purpose of region selection is to identify the region of interest in an image, e.g. the speaker's face in video conferencing. This region is updated automatically by tracking the object's motion. The wavelet-based coder is then applied separately to the foreground and background. Because the size of the region of interest is small, the computation time could be reduced significantly. Adaptive bit allocation is then performed. It makes sure that the video quality of the foreground is always better than that of the background. This is particularly important for unstable networks or low bit rate applications.

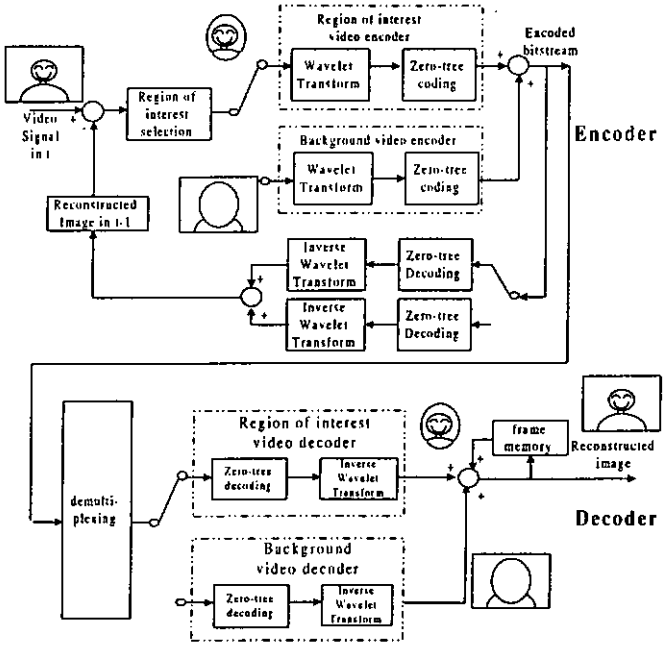


Figure 4.1: The system architecture for the proposed video coder in multipoint video conferencing.

4.3 Region of interest selection

Our proposed system allows the user to specify the region of interest so as to define the foreground and the background initially. If the user does not specify the region, the center region will be used (see Figure 4.2). In any I-frame, the user can change the region of interest which makes the video conferencing interactive. In subsequent frames, the intelligent video conferencing system calculates the difference between the current and the previous frames in order to reduce the temporal redundancy. This difference frame contains information about the motion of the user. In most video conferencing applications, the background is stationary and the difference frame shows only the changed region. Therefore, instead of coding the whole difference frame, only the region that contains large motions needs to be encoded. This reduces the overall bit rate and the encoding time while maintaining a good video quality even under low bit rates. We further propose in this thesis to make an analysis of the histogram of the difference frame,

due to its simplicity and invariance to rotation and translation. The matching criterion is defined as,

$$\sum_{i=0}^{255} [H_{defined}(i) - H_{neighbour}(i)] \tag{4.1}$$

where $H_{defined}(i)$ and $H_{neighbour}(i)$ represent respectively the histogram information in the defined and the neighboring regions. A search is carried out only around the center checking point as shown in Figure 4.2a. By using the histogram [57], instead of the minimum absolute difference as the matching criterion, region tracking with high accuracy can be achieved.

The best match is obtained for a search range when the matching criterion defined in eqn.4.1 is the minimum. If the minimum is found at the center, the procedure stops. Otherwise, further search is conducted around the point where the minimum has just been found. The procedure continues until the winning point becomes a center point of the checking block or when the checking block hits the boundary of the predefined search range. In a practical situation, the interested region can be easily tracked as shown in Figure 4.2b. In summary, the proposed intelligent video conferencing system allows the user to select a region of interest and track the region using a fast algorithm.

Figure 4.2: The proposed searching technique, (a) the region of interest defined by the user and (b) the tracked object in the subsequent frame.

4.4 Adaptive Bit Allocation

Once the region of interest is defined, different video quality for the foreground and the background can be obtained by applying the wavelet transform and the zero-tree coding separately to the region of interest and the background. Therefore, different number of bits can be allocated in different regions so that a good video quality for the foreground can always be guaranteed. Also, the small region of interest can greatly reduce the processing time in the wavelet-based coder. The percentage of bits allocated to the region of interest is specified by the user. Thus the user can control the video quality dynamically.

4.5 Experimental results

The proposed encoder for the multipoint video conferencing system is tested for the 64kbit/sec case. Figure 4.3 shows a comparison between our proposed system and the conventional DCT-based system. The overall performance is shown in Figure 4.4. Although the background has a lower PSNR as compared to the conventional approach, the foreground has a much higher PSNR. Moreover, the subjective performance is much better than the conventional approach and the blocking artifacts are avoided. In fact, the subjective superiority is even more profound at low bit rates. In this case, the blocking artifacts associated with the DCT-based encoders are severe. However, by using our proposed algorithm, the quality in the foreground can still be maintained. Besides, the proposed video conferencing system has an improvement factor of about 2 to 3 times as compared to the case when the wavelet transform is applied to the whole image.

Figure 4.3: Reconstructed frames from (a) our proposed and (b) the DCT-based encoders.

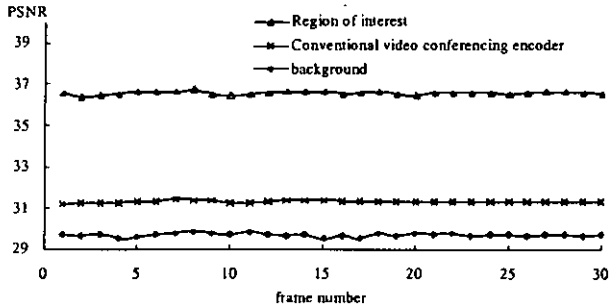


Figure 4.4: A comparison of the PSNR in different frames between our proposed encoder and the DCT-based encoder.

4.6 Conclusions

A region-based video encoder for multipoint video conferencing is proposed. The proposed architecture consists of region of interest selection, wavelet-based encoders for the foreground and the background, motion-based region updating steps and adaptive bit allocation strategy. To make the system interactive, the user can specify the region of interest at any I-frame as well as the relative quality between the foreground and the background. Experimental results confirm that our proposed method produces a good video quality even under low bit rates for real-time video conferencing applications.

Chapter 5
Audio Processing
5.1 A Fast Bit Allocation Algorithm for MPEG Audio Encoder

A fast bit allocation algorithm for the MPEG audio encoder is proposed in this section , which is able to generate an identical MPEG bitstream produced by the standard bit allocation algorithm described in MPEG audio standard. The proposed algorithm employs the bit allocation information of the previous frame as a reference for allocating the restricted bits to each of the 32 subbands in the current frame such that the number of iteration can be significantly reduced. The results show that the performance of the proposed bit allocation algorithm works well at different encoded bitrates.

5.1.1 Introduction of MPEG Audio Coding

The MPEG audio coding scheme[58] is a perceptual audio coding standard and consists of three audio coding algorithms called layer I, II and III. Among these three layers, layer I and layer II have been widely used in multimedia and broadcasting related products[59-60]. Which layer is employed for an application of audio coding is determined by the computational complexity and performance required by the application[61].

Without loss of generality, layer II audio encoder will be discussed in this chapter and its basic structure is shown in Figure 5.1.1. The input audio samples first pass through a subband filter which divides the input signal into 32 equal-width frequency subbands. These subband samples are then uniformly quantized according to the bit allocator, which decides the quantization manners with consideration to the audio quality

and the required bits. In aid of the bit allocator, the psychoacoustic model provides the perceptual resolution, which dynamically calculates the masking threshold for each subband. Any audio signals with levels falling below the corresponding masking thresholds are imperceptible to human ear. Then, the objective of the bit allocator is used to dynamically distribute the available bits amongs the subbands according to the masking threshold provided by the psychoacoustic model. Theoretically, by coding at the demanded bitrate, a perceptually lossless quality of audio signals can be achieved. If the demanded bitrate is higher than the available bitrate, the bit allocator tries to minimize the total noise-to-mask ratio(NMR) over the frame with the constraint that the number of bits required does not exceed the number of bits available for that frame. After the bit allocation process, the subband samples are then scaled, quantized according to the bit allocation information, and formatted into an encoded MPEG bitstream together with a header, bit allocation and scaling information. In the standard bit allocation procedure, over 100 iterations on average are needed which is computational intensive. Motivated by this, a fast and efficient algorithm for bit allocation scheme is proposed in this chapter. It is found that the proposed fast algorithm is able to strengthen the conventional bit allocation algorithm at different encoded bitrates.

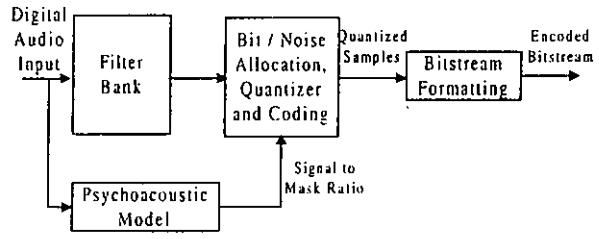


Figure 5.1.1 Basic structure of the audio encoder.

5.1.2 Bit Allocation procedure

The MPEG coding scheme achieves the compression by placing the quantization noise in the frequency subbands where the ear is least sensitive. The psychoacoustic model determines from the maximum noise level of the input audio which would just be perceptible (masking level) for each of the subbands. The quantitative sketch of figure 5.1.2 gives a few more details about the masking threshold. Within a critical band, tones below this threshold are masked. As the amount of the quantization noise is directly related to the number of bits used by the quantizer, the bit allocation algorithm assigns the available bits in a manner which minimizes the audible distortion. The psychoacoustic model described in the standard returns the Signal to Mask Ratio (SMR) for each subband, which is defined as the difference in dB between the level of masker and the minimum masking threshold within the critical band. Assuming an m-bit quantizaion of an audio signal, within the critical band the quantization noise will not be audible as long as its signal-to-noise ratio(SNR) is higher than its SMR. The bit allocation algorithm computes the Noise-to-Mask Ratio(NMR) from the SMR using the following expression.

$$NMR(m)=SMR-SNR(m) \quad \text{(in dB)} \tag{5.1.1}$$

NMR(m) describes the difference in dB between the SMR and the SNR to be expected from an m-bit quantization. The NMR value is also the difference (in dB) between the level of quantization noise and the level where a distortion may just become audible in a given subband. Within a critical band, coding noise will not be audible as long as NMR(m) is negative.

The bit allocator looks at both the outputs samples from the filterbank and the SMR from the psychoacoustic model, and adjusts the bit allocation in order simultaneously to meet both the bitrate requirements and the masking requirements. Bit allocation algorithm recommended by the MPEG standard involves a number of iterations where, in each iteration, the number of quantizing levels of the subband with the largest NMR is decreased as long as the number of bits used does not exceed the total number of bits available for the frame. The NMR in dB for each subband is obtained from equation (5.1.1). Each incremental bit assigned to a subband will incur 36 bits as there are a total of 36 time samples within each subband. As a result, the iteration will stop once the number of bits available for coding is smaller than 36.

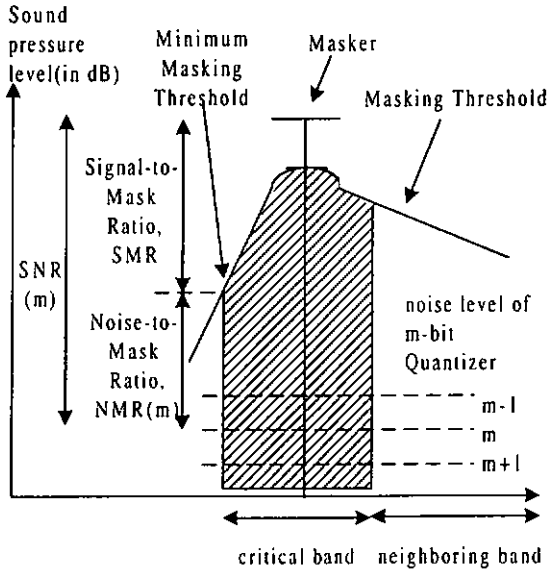


Figure 5.1.2 Masking threshold and signal-to-mask ratio(SMR).

The procedure of allocating the bits involves a large number of iterations. Therefore, it demands an enormous amount of computational steps. Recently, a fast bit allocation algorithm has been proposed[62], which seeks to reduce the number of the iteration by computing the demand bitrates, which is defined as the bit allocation needed to give a zero or just negative value of NMR for every subband. This demand bitrate is

then substracted from the total available bit rate as constrained by the channel to obtain the available bitrate for allocation. If the available bitrate is greater than or equal to 36 bits per frame, the available bits will be allocated according to the standard iterative procedure until all the bits are exhausted. However, if the available bit rate is less than zero, the subband with the smallest NMR is identified and the number of quantizing steps allocated to it is reduced and, consequently, the NMR of the subband is also increased. This iterative procedure is carried out until the available bitrate becomes positive. The number of iterations of this algorithm is significantly reduced when the allowable bitrate is very close to the demanded bitrate. However, it breaks down at very low bitrates because of the large number of iterations needed to reduce the number of bits allocated to each band. In this chapter, we propose a fast bit allocation algorithm, which works well at a different bitrates and has simple computational complexity.

5.1.3 Proposed fast bit allocation

The computation of bit allocation in the MPEG is an iterative procedure. Bits are allocated one at a time to the subband with the maximum NMR, and that subband's NMR is reduced accordingly. Furthermore, for each bit allocated, the NMR values of all subbands in all channels are scanned to select the subband with the maximum NMR value. This is a very time consuming process. Motivated by this, a fast bit allocation algorithm which requires a few numbers of iterations is proposed to reduce the computation complexity. In general audio signals, the current frame is highly correlated with the previous frame. Due to this correlation property, the bit allocation information of current frame is very similar to the previous one. Figure 5.1.3(a), 3(b) and 3(c) show

the similarity of bit allocation between frame *N-1* and *N* in 128kbit/s, 96kbit/s and 64kbit/s respectively.

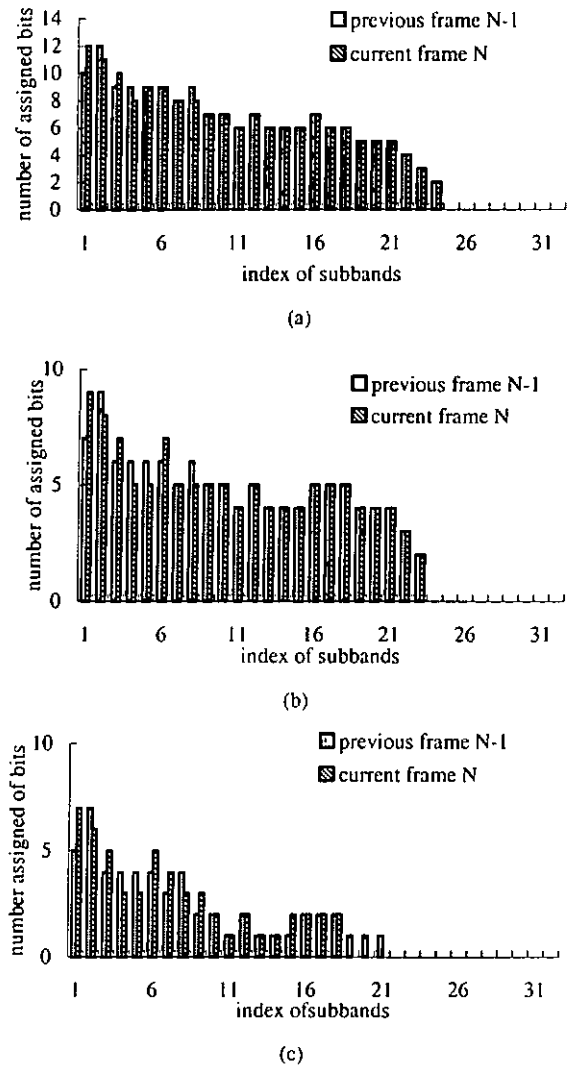


Figure 5.1.3 The number of bits assigned to each subband at different encoding bitrates (a) 128kbit/s, (b) 96kbit/s and (c) 64kbit/s.

The correlation behaviour of the bit allocation information provides a helpful guideline to speed up the iterative process of the bit allocation algorithm. Consequently, the number of bits assigned to each subband of current frame can be predicted from that of the neighbour frame. In other words, the bit allocation information of previous frame is employed to form a good initial estimate for the current frame. Then the iterative steps

of bit allocation are performed to either allocate the bits to or deallocate the bits from the appropriated subbands. These two possible cases are shown in Figure 5.1.4. If the available bits are greater than or equal to 36 bits per frame, the bit allocation algorithm now attempts to minimize the maximum NMR of all subbands by assigning the remaining bits to the subband samples. The algorithm computes the NMR for each subband and finds the subband with the maximum NMR. The bit allocation for that subband is increased one level and the number of additional bits required is subtracted from the available bits. The process is repeated until all the available bits have been used. This is same as the standard bit allocation process suggested in the MPEG audio standard. If the available bits are less than 36 bits per frame, the subband with the minimum NMR is determined and the number of bits allocated to it is reduced by one level. As a result, the NMR of the subband is increased. This iterative process is performed until the available bits become positive. This algorithm indicates that if the bit allocation information of the current frame can be predicted from the previous frame, it is not necessary to waste much computation time to perform the iteration.

However, the number of bits assigned to each subband by the proposed bit allocation algorithm may not be identical to that of the standard bit allocation algorithm suggested in the MPEG audio standard. A reallocation of bits among all of subbands is required in order to produce the same bit allocation information of the standard algorithm, as depicted in Figure 5.1.4. Let NMR_i^n denote the NMR of subband i in the iteration step n. Assume that NMR_k^n and NMR_h^n are minimum NMR and maximum NMR in step n. In order to obtain the optimum bit allocation, the number of bits allocated to subband k is reduced by 36 bits which are reallocated to the subband h. The process is repeated until the

minimum NMR is exchanged with the maximum NMR in the next iteration steps. In order words, the reallocation process is stoped when NMR_k^{n+1} and NMR_h^{n+1} are exchanged to maximum NMR and minimum NMR in step $n+1$, respectively. The convergence of the reallocation process ensures that the resulting encoder is able to produce a MPEG compliant bitstream which is almost identical to the bitstream produced by the reference MPEG encoder suggested in the standard.

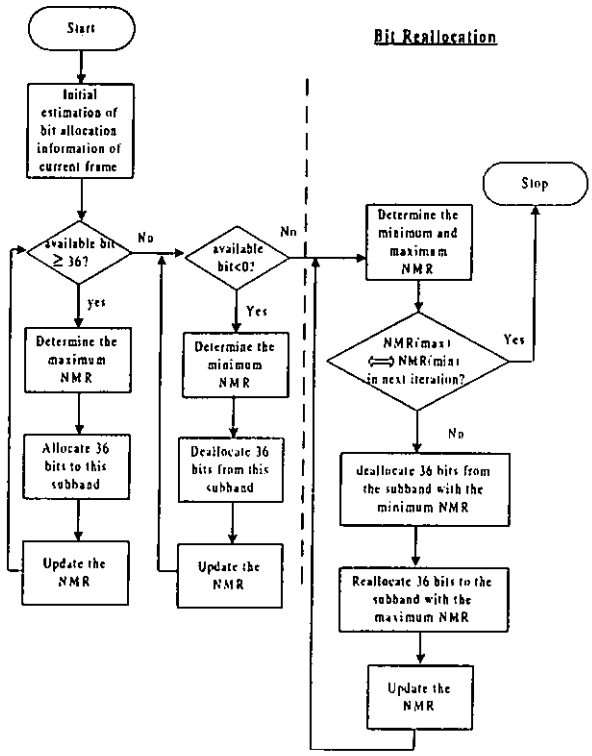


Figure 5.1.4 The flowchart of the bit reallocation process.

5.1.4 Simulation Results and Discussion

To compare the performance of the proposed bit allocation algorithm with the conventional algorithms, a number of audio sequences were coded at different bitrates 128kbit/s, 96kbit/s and 64kbit/s per monophonic channel. The proposed algorithm is compared with the standard bit allocation algorithm described in the MPEG audio

standard[58] and the Teh's bit allocation algorithm[62]. The simulation results are shown with an average number of iterations and a speed-up ratio with reference to that of the standard bit allocation algorithm, which are summarized in Table 5.1.1 and Table 5.1.2. From these tables, our algorithm has the best performance among all schemes at different bitrates. At bitrates of 96kbit/s and 128kbit/s, the encoded bitrates are greater than the demanded bitrate. The Teh's algorithm works well, but the proposed algorithm has further speed-up, about 4 to 6 as compared with the Teh's algorithm. The results indicate that the proposed algorithm is able to exploit the correlation of neighbour frame to alleviate the complexity of bit allocation process in audio coding. Figure 5.1.5 shows the number of iterations taken by both algorithms against the frame number at 64kbit/s. It shows that the Teh's algorithm breaks down at very low bitrates. At this bitrate, the Teh's algorithm performs much worse than the standard algorithm. This is because of the large number of iterations needed to reduce the number of bits allocated to each subband. Again, the proposed algorithm demonstrates a better performance, as shown in Table 5.1.1, Table 5.1.2 and Figure 5.1.5. This further supports the advantage of the proposed bit allocation algorithm.

Table 5.1.1: Comparison of performance in terms of number of average iterations for different bit allocation algorithms.

Bit rate	Standard algorithm[1]	Teh's algorithm[5]	Proposed algorithm
64kbit/s	53	63.8	7.4
96kbit/s	105	20.6	5.2
128kbit/s	163	25.2	4.4

Table 5.1.2: Speed up ratio of different bit allocation algorithms as compared with the standard algorithm.

Bit rate	Teh's algorithm[5]	Proposed algorithm
64kbit/s	0.83	7.16
96kbit/s	5.10	20.19
128kbit/s	6.47	37.05

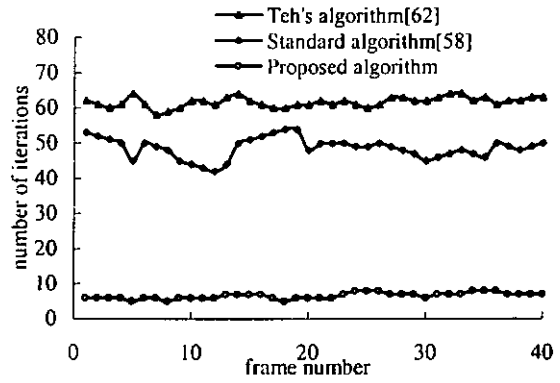


Figure 5.1.5 Iterations against frame number at 64kbit/s.

Figure 5.1.6 analyzes the average number of iterations in our proposed algorithm. About 19%, 9% and 4% of iterations is used for bit reallocation at 64kbit/s, 96kbit/s and 128kbit/s respectively, which ensures the bit allocation information produced by the proposed algorithm is identical to that of algorithm described in the MPEG audio standard but with a reduced complexity.

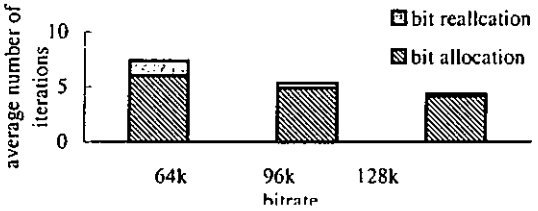


Figure 5.1.6 Bit reallocation against bit allocation in the proposed bit allocation algorithm.

5.1.5 Conclusion

A fast bit allocation algorithm for the MPEG encoder has been proposed in this chapter. Since the audio information is highly correlated from frame to frame, it is beneficial to use the previous frame as a reference to speed up the iteration process. The proposed scheme can achieve significant reduction of iterations in bit allocation process at different encoded bitrates. Also, the reallocation process ensures that our algorithm

provides the quality of the decoded sequences which is almost identical to that of the algorithm suggested in MPEG audio standard. Experimental results show that the proposed algorithm is more efficient in bit allocation than other algorithms.

5.2 A Constrained Optimisation Approach to Speech Signal Recovery

In this section, we address a problem of speech enhancement, which is to recover a speech source from a mixture of its delayed versions and additive noise. By using the constrained optimisation technique, the second order statistics based algorithm is developed. The new proposed algorithm requires no strong limitations to the speech signal and the noise. Simulation results show that our algorithm achieves a better performance as compared to other algorithms.

5.2.1 Introduction

Speech enhancement has been active in speech signal processing. The objective is to extract a single speech source signal from its delayed versions and in noisy environment. Depending upon the amount and the type of noise, and the strength of echoes existing in the environment, the resulting speech signals could vary substantially. The quality of the speech may range from being slightly degraded to being annoying to listeners, and in the worst case it could be totally unintelligible. It is very necessary to recover the speech signal from the distortion.

A number of approaches to signal recovery have been proposed [35-38]. These approaches make use of the output second-order statistics [35-36] or the output higher-order statistics [37-38]. They are basically the least squares solutions. However, if the unknown parameters in an algorithm have inherent non-linear relations, they are usually

assumed to be independent from each other to apply the linear least squares technique. A larger estimation error is inevitably generated, although additional post-processing step may reduce the error. To accommodate practical applications and achieve accurate estimation, we will design our algorithm with the following features, (1) no strong limitations to the source and noise except for all signals being stationary, and (2) considering the inherent non-linear relation among the unknown parameters while deriving our new algorithm to avoid the additional post-processing step.

5.2.2 Problem and Assumptions

Let us firstly consider a linear time invariant (LTI) system with the following model.

$$\begin{aligned} y_1(k) &= A(z^{-1})s(k) + T(z^{-1})n(k) \\ y_2(k) &= C(z^{-1})s(k) + n(k) \end{aligned} \tag{5.2.1}$$

where $A(z^{-1}) = a(0) + a(1)z^{-1} + \Lambda + a(p)z^{-p}$,

$C(z^{-1}) = c(0) + c(1)z^{-1} + \Lambda + c(r)z^{-r}$,

$T(z^{-1}) = t(0) + t(1)z^{-1} + \Lambda + t(l)z^{-l}$.

z^{-1} is a shift operator, i.e., $z^{-1}s(k) = s(k-1)$. The received signals $y_1(k)$ and $y_2(k)$ are the outputs of the LTI systems with the same input $s(k)$, $n(k)$ is the received noise from a receiver. Because the two receivers are in the same background, the two received noises are highly related. We use a linear time-invariant operator $T(z^{-1})$ to link the received noises.

To simplify the problem, we assume that:

All signals are sampled wide-sense stationary random processes with zero-mean.

$A(z^{-1})$ and $C(z^{-1})$ are relatively prime, and $a(0)=1$.

5.2.3 Algorithm Development

Unconstrained LS solution To accommodate practical environment, we will make use of the second-order statistics of the outputs for the parameter estimation. The cross- and auto-correlation of $y_i(k)$ and $y_j(k)$ for lag τ are represented by:

$$r_{y_i y_j}(\tau) = E[y_i(k)\, y_j(k+\tau)] \qquad i, j=1,2. \tag{5.2.2}$$

Substituting eqn.5.2.1 into eqn.5.2.2 and applying the Fourier transform, the following results can be obtained:

$$\begin{aligned} &A(\omega)P_{y_2y_1}(\omega) - A(\omega)T^*(\omega)P_{y_2y_2}(\omega) \\ &= C(\omega)P_{y_1y_1}(\omega) - T^*(\omega)C(\omega)P_{y_1y_2}(\omega) \end{aligned} \tag{5.2.3}$$

where $P_{y_i y_j}(\omega)$ $(i, j=1, 2)$ is the power spectrum of the joint random processes y_i and y_j, $P_{ss}(\omega)$ and $P_{nn}(\omega)$ are the power spectrums of the source signal and the additive noise, respectively. Let

$$\begin{aligned} A_l(\omega) &= e^{-j\omega l} A(\omega)T^*(\omega) \\ C_l(\omega) &= e^{-j\omega l} C(\omega)T^*(\omega) \end{aligned} \tag{5.2.4}$$

where $A_l(\omega) = a_l(0) + a_l(1)e^{-j\omega} + \Lambda + a_l(p+l)e^{-j(p+l)\omega}$ and $C_l(\omega) = c_l(0) + c_l(1)e^{-j\omega} + \Lambda + c_l(r+l)e^{-j(r+l)\omega}$.

Substituting eqn.5.2.4 into eqn.5.2.3 and taking the inverse Fourier transform, we have

$$\begin{aligned} &\sum_{i=1}^{p} a(i) r_{y_2y_1}(k-i) - \sum_{i=0}^{p+l} a_l(i) r_{y_2y_2}(k+l-i) \\ &- \sum_{i=0}^{r} c(i) r_{y_1y_1}(k-i) + \sum_{i=0}^{r+l} c_l(i) r_{y_1y_2}(k+l-i) = -r_{y_2y_1}(k) \end{aligned} \tag{5.2.5}$$

Eqn.5.2.5 may be viewed as a linear equation with $(2p+2l+2r+3)$ unknowns involving $\{a(k)\ k=1,2,\ldots,p\}$, $\{c(k)\ k=0,1,2,\ldots,r\}$, $\{a_l(k)\ k=0,1,2,\ldots,p+l\}$ and $\{c_l(k)\ k=0,1,2,\ldots,r+l\}$ if unknowns $\{a_l(k)\}$ and $\{c_l(k)\}$ are assumed to be independent of $\{a(k)\}$ and $\{c(k)\}$, respectively. By selecting some values for k in series, a set of overdetermined equations

is formed. For instance, let k range from m_1 to m_2, eqn.5.2.5 can be expressed in matrix form:

$$R\theta = r \tag{5.2.6}$$

where

$$\theta = [a(1)...a(p) \quad c(0)...c(r) \quad a_t(0)...a_t(p+l) \quad c_t(0)...c_t(r=l)]^T,$$

$$r = \left[r_{y_2y_1}(m_1) \quad r_{y_2y_1}(m_1+1) \quad \ldots \quad r_{y_2y_1}(m_2)\right]^T.$$

$$R = \begin{bmatrix} -r_{y_2y_1}(ml-1) & \ldots & -r_{y_2y_1}(ml-p) & r_{y_1y_1}(ml) & \ldots & r_{y_1y_1}(ml-r) \\ -r_{y_2y_1}(ml) & \ldots & -r_{y_2y_1}(ml+1-p) & r_{y_1y_1}(ml+1) & \ldots & r_{y_1y_1}(ml+1-r) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ -r_{y_2y_1}(m2-1) & \ldots & -r_{y_2y_1}(m2-p) & r_{y_1y_1}(m2) & \ldots & r_{y_1y_1}(m2-r) \\ r_{y_2y_2}(ml+l) & \ldots & r_{y_2y_2}(ml-p) & -r_{y_1y_2}(ml+l) & \ldots & -r_{y_1y_2}(ml-r) \\ r_{y_2y_2}(ml+l+1) & \ldots & r_{y_2y_2}(ml+1-p) & -r_{y_1y_2}(ml+l+1) & \ldots & -r_{y_1y_2}(ml+1-r) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{y_2y_2}(m2+l) & \ldots & r_{y_2y_2}(m2-p) & -r_{y_1y_2}(m2+l) & \ldots & -r_{y_1y_2}(m2-r) \end{bmatrix}$$

The corresponding LS solution is

$$\theta = (R^T R)^{-1} Rr \tag{5.2.7}$$

Note that it is not a final solution for each coefficient. The final solution is usually obtained by synthesising the LS estimates that are related but not are the coefficients in $A(z^{-1})$ and $C(z^{-1})$. This processing is called as post-processing. Obviously, the LS solutions may be far away the real parameters due to ignoring the inherent non-linear relations between $\{a(k)\}$ and $\{a_t(k)\}$, and $\{c(k)\}$ and $\{c_t(k)\}$. To avoid additional post-processing step and reduce estimation error, let us consider these relations while deriving our algorithm in the next section.

Constrained optimal solution Let $e = R\theta - r$. Set the goal attainment J as

$$J = e^T e \tag{5.2.8}$$

Minimising J produces the above unconstrained LS solution, which ignores the inherent non-linear relations of the unknowns. To obtain reasonable solution, let us consider these relations together. From eqn.5.2.4, it is very easy to attain the following $(p+r+l+1)$ equations.

$$\sum_{i=0}^{k} a_t(i)c(k-i) = \sum_{i=0}^{k} a(i)c_t(k-i) \qquad \text{for } k = 0,1,\ldots, p+r+l. \tag{5.2.9}$$

A constrained optimisation problem can then be stated as follows

$$\begin{aligned} &\text{Min } J \\ &\text{Subject to } \sum_{i=0}^{k} a_t(i)c(k-i) = \sum_{i=0}^{k} a(i)c_t(k-i) \\ &\qquad\qquad k = 0,1,\ldots\ldots, p+r+l \end{aligned} \tag{5.2.10}$$

By using the Gauss-Newton method, we can obtain the optimal solutions for $\{a(k), k=1,2,\ldots,p\}$, $\{c(k), k=0,1,2,\ldots,r\}$, $\{a_t(k), k=0,1,\ldots,p+l\}$ and $\{c_t(k), k=0,1,\ldots,r+l\}$. We note that this is not the only algorithm for solving the criterion equations. Iterative gradient-based algorithms such like the steepest-descent or the Newton-Raphson may also be applied here.

Signal reconstruction Provided that the unknown channel parameters have been identified by the above method denoted by $\hat{A}(z^{-1})$, $\hat{C}(z^{-1})$ and $\hat{T}(z^{-1})$, respectively, and that all the roots of $\hat{A}(z^{-1}) - \hat{C}(z^{-1})\hat{T}(z^{-1}) = 0$ are inside the unit circle, the estimated signal can be expressed by

$$\hat{s}(k) = \frac{1}{\hat{A}(z^{-1}) - \hat{C}(z^{-1})\hat{T}(z^{-1})}[y_1(k) - \hat{T}(z^{-1})y_2(k)] \tag{5.2.11}$$

Obviously, if $\hat{A}(z^{-1})$, $\hat{C}(z^{-1})$ and $\hat{T}(z^{-1})$ equal to $A(z^{-1})$, $C(z^{-1})$ and $T(z^{-1})$, respectively, the estimated source exactly equals the real one.

5.2.4. Simulation Results

Extensive simulations have been carried out to compare the proposed constrained optimisation algorithm (COP) with the higher order statistics-based (HOS) [64] and the unconstrained least square algorithm (ULS). For each test case, the coefficient vectors in model (5.2.1) are $[a(0)\ a(1)\ a(2)]=[1\ 0.2\ 0.1]$, $[c(0)\ c(1)]=[0.8\ 0.2]$ and $[t(0)\ t(1)]=[0.7\ 0.15]$. We define the signal-to-noise ratio as $SNR = 10\log(\| s(\cdot)\|_2 / \| n(\cdot)\|_2)$ (dB) and define the mean squared error (*MSE*) as $MSE = [(h-\hat{h})^T(h-\hat{h})/h^T h]^{1/2}$, where $\hat{h}$ is a vector that consists of all estimated coefficients and h is a vector that consists of real values.

Experiment 1: Parameter estimation

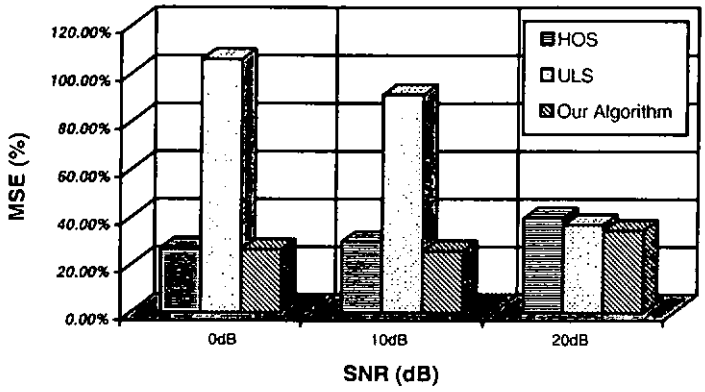
We add a digital signal (the data length=2000) as the source signal, which was generated by

$$15[\sin(0.1\pi k)+\sin(0.3\pi k)+\sin(0.6\pi k)] \tag{5.2.12}$$

Two cases for the additive noise are considered.

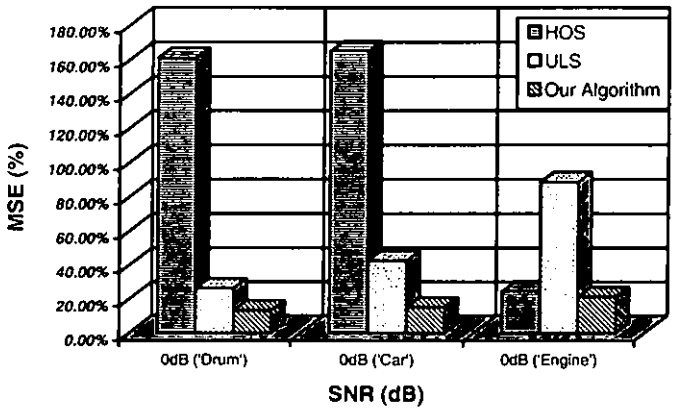
Gaussian noise: In this case, only Gaussian noise is added to model (5.2.1). Figure 5.2.1 shows the percentages of *MSEs* of the estimates for each algorithm at SNR 0dB, 10dB and 20dB, respectively. From this chart, it is easy to observe that the percentages of *MSEs* for both the HOS and the COP algorithms have little change when SNR changes from 0dB to 20dB, whereas the ULS algorithm has a distinct change. This demonstrates that although the COP algorithm is not designed for the Gaussian noise, it can achieve similar accuracy with the HOS algorithm even if the SNR is lower.

Figure 5.2.1: MSEs of parameter estimates with Gaussian white noise (data length=2000)



Real noises: As we known, it is difficult to verify that the additive noise is of Gaussian property or not in practice. We thus need to test the suitability of the algorithms to some real noises. For this goal, we test three real noises using the same model. They are "drum", "car" and "engine". Figure 5.2.2 shows the corresponding results for each approach only at SNR 0dB. From this chart, we can see that only the COP algorithm can adapt these noises with better accuracy, whereas the HOS algorithm produces very large estimation errors in the cases of "drum" and "car". It is not surprising, because the COP algorithm is not limited to any kinds of noises but the HOS algorithm is available only for Gaussian noise. "Engine" may be close to Gaussian noise, the estimation accuracy thus becomes better for the HOS algorithm.

Figure 5.2.2: MSEs of the estimated parameters with real noises (data length=2000)



Experiment 2: Signal recovery

In this experiment, let us see the effects of signal recovery using the COP algorithm. The reconstructed signals by using the other two algorithms are not involved in these experiments. On one hand, the estimated parameters by these two algorithms have larger errors compared to our algorithm. The corresponding reconstructed signals must be worse than using our algorithm. On the other hand, even if parameter estimation is performed, in many cases the reconstructed signals can not be obtained from the two algorithms, as the third assumption is not satisfied. We add the signal to be used in experiment 1 and speech signal to model (5.2.1) with real noises at SNR 0dB, respectively. After parameter estimation, the signals to be reconstructed are shown in Figures 5.2.3 and 5.2.4. From Figures 5.2.3, we can see that the estimated signal is very close to the original one with very small errors. But Figures 5.2.4 illustrates larger errors between the estimated signal and the real speech signal. The reason is that speech signal, in most cases, is not stationary, whereas our algorithm requires the source signal should be stationary.

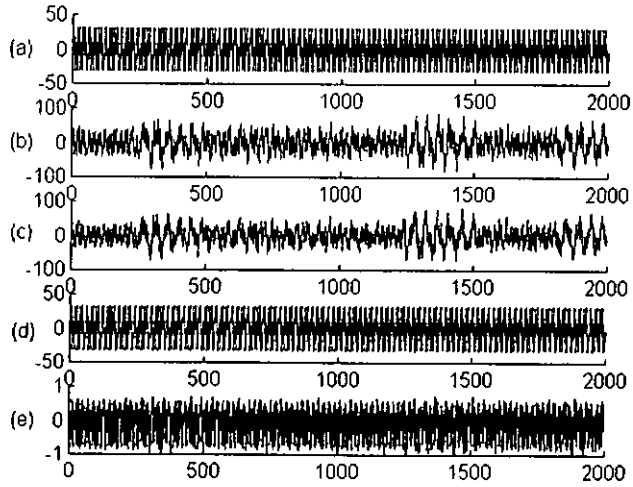


Figure 5.2.3: Signal recovery with "drum" noise at SNR 0dB via our algorithm (a) original source signal; (b) and (c) two received signals; (d) estimated signal; (e) the error between (a) and (d).

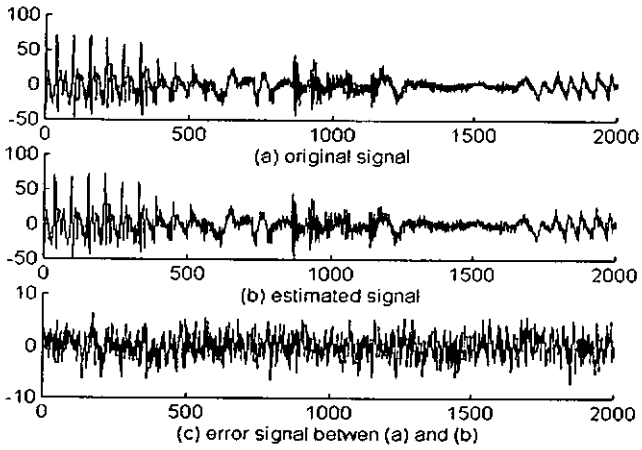


Figure 5.2.4 Speech enhancement with "engine" noise at SNR 0dB via our algorithm (a) original speech signal; (b) estimated signal; (c) the error between (a) and (c).

5.2.5 Conclusions

A new constrained optimisation algorithm for FIR channel identification and speech enhancement has been proposed in this section. Because of using the second order statistics of outputs, there is no strong limitations to the source signal and additive noise. In addition, by employing the constrained optimisation technique our algorithm avoids the additional

post-processing step that usually happens in other algorithms. Simulation results have illustrated the correctness of the coefficients using our new algorithm. By comparing our algorithm with the ULS and the HOS algorithms, we can see that our algorithm performs the parameter estimation and recoveries the original speech signal with a better performance under the practical noises even if the SNR is 0dB.

Chapter 6

Conclusion and Possible future work

6.1 Conclusion of the present work

In this thesis, we give results for an investigation on scalable video and audio techniques for video conferencing. The scalable techniques are very powerful for video and audio coding which can improve the reconstruction video and audio quality as well as the computational speed of existing algorithms. According to the nature of the techniques, our investigation can be divided into two parts, that is, the video processing and the audio processing.

The investigation on the video processing techniques based on the video conferencing participants behaviour has been presented in Chapter 3 and Chapter 4. In these two chapters, a possible adaptive solutions to solve the problems of low bitrate video conferencing by assigning more bits to the active participant. These techniques include approaches that take into account of active region and background or inactive region in a video conferencing system. In chapter 3, the quality degradation in a transcoding process is addressed. It has been shown that during the transcoding process, the re-encoding error is introduced. Besides, it suffers from the intrinsic problem of double encoding in transcoding. A study of the behaviour of active speaker has indicated that various adaptive schemes, such as direct summation of DCT coefficients, DCT-domain buffer updating for motion-compensated macroblock, frame skipping criterion, can be used effectively to improve the video quality and coding efficiency. A comparison between the present approach and the conventional transcoder is also performed. It has a speed-up of about 7 times faster than that of the conventional

transcoder for the "Salesman" sequence,say for example. This is because the probability a macroblock being coded without motion compensation happens more frequently in typical sequences, and this type of macroblocks should not introduce any re-encoding error due to the direct summation of the DCT coefficients. The average PSNR performance of the macroblock coded without motion compensation in our proposed transcoder is significantly better than that of the conventional transcoder. Thus, we can achieve significant computational savings while maintaining a good video quality on these macroblocks. On the other hand, our proposed transcoder also shows an improvement with respect to the motion-compensated macroblock. Furthermore, the cache system in the transcoder can reduce the computational burden of re-encoding the motion-compensated macroblocks. All these advantages combined gives rise to a significant computational saving as well as quality improvement. These demonstrate the effectiveness of the proposed frame-skipping transcoder. Besides, the performance of the proposed video combiner for continuous presence multipoint video conferencing system as compared to the conventional pixel-domain combiner with dynamic frame-skipping is evaluated. Since the active participants need higher frame-rates to produce an acceptable video quality while the inactive participants only need lower frame-rates to produce an acceptable quality, we use a dynamic frame allocation in both the proposed video combiner and the PDCOMB-DFS to distribute the encoded frames into each sub-sequence according to the motion activities. Inevitably, this improvement is made by sacrificing a certain amount of quality of the motion inactive periods. The active period is transcoded more frequently following the motion activities of the sub-sequence, therefore the videos displayed on the receiver are smoother. It has been shown that the

conventional PDCOMB-DFS suffers loses due to the double-encoding aspect while the proposed video combiner offers a much better quality as compared to the PDCOMB-DFS. The gain can be as high as 1.5-2.0 dB for both the active and non-active periods due to the high efficiency of our proposed frame-skipping transcoder. It can be seen that the video quality of the active participant with the proposed video combiner is much better than that with the PDCOMB-DFS. For using the dynamic frame allocation which is based on the motion activities, the degradation of video quality of the inactive participants is not very visible. By using the proposed video combiner, the PSNR's for all conference participants are much improved as compared to the PDCOMB-DFS during both the active and non-active periods. In a practical multipoint video conferencing system, active participants are given most attention. An improvement of the video quality of the active participants is particularly important and we have shown that the proposed video combiner can achieve a significant improvement.

Chapter 4 gives a new direction for the video conferencing system. In the conventional approach it is always to give focus on the block-based video coding. In fact, a good performance can be achieved with a large bandwidth. However, under low bit rates, the DCT-based encoder exhibits visually annoying blocking artifacts. A region based video coder which produces different video qualities in the foreground and background is proposed in this thesis. The coder makes use of the wavelet transform. After using the wavelet transform, a high quality video with the absence of blocking artifacts can be achieved. Although a wavelet-based coder can achieve a good quality, its computational complexity is a complicated problem. A way to speed up the computation is to explore the fact that the various regions in an image are not of equal importance.

The proposed video coder is based on the adaptive region-based updating technique by which the video is updated according to the motion activity. It has been found that the proposed video coder allows a high quality video in the region of interest while it also reduces the overall bit rate and computation time. Since a user might be in fast motion when active, a simple and fast object tracking technique is proposed to locate the region of interest. This approach produces a good video quality even under low bit rates. Besides, in order to make the system interactive, a user can specify the region of interest at any time instant as well as the relative quality between the foreground and the background.

An investigation on audio processing techniques for an improvement of the audio quality in video conferencing has been presented in Chapter 5. In Chapter5, a brief introduction of MPEG Audio Coding is given and the scheme using over 100 iterations in the bit allocation process is shown. The conventional approach in bit allocation process is reviewed and the problem of Teh's algorithm is studied. In the Teh's algorithm, the number of iterations is greatly reduced when the demanded bitrate is similar to the encoded bitrate. However, it is proved that the algorithm fails in low bitrate encoding such as 64kbit/sec. By using the previous frame bit allocation information as a reference, our proposed algorithm can finish the iteration process in bit allocation within only a few iterations. With the help of the proposed bit re-allocation process, the bits can be guaranteed to reallocate to minimize the most noticeable quantization noise. As a result, almost identical bitstreams can be produced compared with the conventional approach while only a few iterations are required. In the second half of this chapter, we address the problem of speech enhancement, which is to recover a speech source from a mixture of

its delayed versions and some additive noises. Depending upon the amount and the type of noises, and the strength of echoes existing in the environment, the resulting speech signals could vary substantially. The quality of the speech may range from being slightly degraded to annoying to listeners, and in the worst case it could be totally unintelligible. It is very necessary to recover the speech signal from the distortion. By using the constrained optimisation technique presented in chapter 5, a second order statistics based algorithm is developed. The newly proposed algorithm requires no strong limitations to the speech signal and the noise. The algorithm is realized in a practical environment and an outstanding performance in speech enhancement can be achieved. However, the computational time and the stability also need to be improved such that the proposed signal recovery system can be used practically in our video conferencing system under a noisy environment.

In conclusion, as the technologies in video application such as video conferencing and video broadcasting become more popular. It is important to apply some scalable techniques to videos to reduce the huge size for storage and transmission for different applications. In present work, we have shown that for designing a video conferencing system, the scalable and interactive techniques have played an important role as compared to conventional approaches. Results of our of investigation on using the adaptive techniques and interactive functionality in the video conferencing system in this thesis offers new understanding of the video conferencing system. Besides, the audio processing such as high quality MPEG coding scheme which can be applied in sending voice mail and for speech enhancement can be achieved even under an noisy

environment. After all, we sincerely believe that the results obtained in this work are significant to an efficient realization of a modern video conferencing system.

6.2 Future Work

The continuous researches in theoretical and algorithmic adaptive techniques have greatly improved the performance, the practicality and the robustness of the video conferencing system. These adaptive techniques can really upgrade the existing video conferencing system. There are a lot of successful applications appearing in the recent literature. The emphasis in this thesis is also based on the enhancement of the existing video conferencing system. However, with the very strict demands for video conferencing applications which change everyday in terms of quality, speed and scalability, we give some opinions for possible future development of our related studies as shown below.

In the video transcoding, we have proposed the dynamic frame-skipping transcoder which can reduce the quality degradation significantly and reduce the computational complexity. In order to achieve a higher transcoding ratio, combining the requantization transcoder and the frame-skipping transcoder will be our future goal.

Besides the conventional approach in the video conferencing, we have proposed a region-based wavelet coder which provides a better quality in the region of interest and upgrades the background quality if bits are available. The main contribution in this video conferencing system is not only that it can avoid the blocking artifacts and achieve the better quality in the interested region, but it also can provide the interactive functionality to assign the bit ratio to be spent in the foreground and background and can change the region of interest in any time instant. However, the wavelet-based coder has a higher

computational complexity as compared to the conventional video coder. In order to further reduce the computational complexity in the proposed video coder, a fast algorithm and adaptive scheme will be our future work.

In the audio processing, we have proposed a fast bit allocation algorithm in MPEG Audio and speech enhancement. The proposed fast bit allocation algorithm can achieve a significant improvement in the bit allocation process. In order to have a further improvement in terms of the computational complexity, we may consider some adaptive schemes and fast algorithms in the psychoacoustic model and subband filtering in the future. In the speech enhancement, we have addressed a problem of speech enhancement. This is to recover a speech source from a mixture of its delayed versions and an additive noise. Depending upon the amount and the type of noise, and the strength of echoes existing in the environment, the resulting speech signals could vary substantially. The quality of the speech may range from being slightly degraded to being annoying to listeners, and in the worst case it could be totally unintelligible. It is very necessary to recover the speech signal from the distortion. By using the constrained optimisation technique, the second order statistics based algorithm is developed in chapter 5. The new proposed algorithm requires no strong limitations to the speech signal and the noise. Although the speech recovery is good in both theoretical and practical situations, however, the performance is not very stable under a practical environment in sometimes. Some feedback or recursive techniques may be consider such that the separation performance can be improved.

References:

[1] Jeongnam Youn, Ming-Ting Sun and Chia-Wen Lin, "Motion vector refinement for high-performance transcoding," IEEE Transactions on Multimedia, vol. 1, pp. 30-40, March 1999.

[2] Yui-Lam Chan and Wan-Chi Siu, "Variable Temporal Length 3-D Discrete Cosine Transform Coding," IEEE Transactions on Image Processing, vol. 6, no.5, pp.758-763, May 1997, U.S.A.

[3] Video Coding for Low Bitrate Communication, ITU-T Recommendation H.263, May 1997.

[4] ISO/IEC 11172-2, "Information Technology -- Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1,5 Mbit/s -- Part 2: Video," 1993

[5] ISO/IEC 13818-2, "Information Technology -- Generic Coding of Moving Pictures and Associated Audio Information: Video," 1996.

[6] Shaw-Min Lei, Ting-Chung Chen, and Ming-Ting. Sun, "Video bridging based on H.261 standard," IEEE Transactions on Circuit Theory, vol. 4, pp. 425-437, August 1994.

[7] Ming-Ting Sun, Tzong-Der Wu, and Jenq-Neng Hwang, "Dynamic bit allocation in video combining for multipoint conferencing," IEEE Trans. on Circuits and Systems – II: Analog and Digital Signal Processing, vol. 45, no. 5, May 1998.

[8] ITU-T Study Group XV, Recommendation H.261, "Video codecs for audiovisual services at $p \times 64$ kb/s," May 1992.

[9] Ming-Ting Sun, Alexander C. Loui, and Ting-Chung Chen, "A coded-domain video combiner for multipoint continuous presence video conferencing," IEEE Transactions on Circuits and Systems for Video Technology, vol. 7, no. 6, pp. 855-863, December 1997.

[10] H. Sun, W. Kwok and J.W. Zdepski, "Architectures for MPEG compressed bitstream scaling," IEEE Transactions on Circuits and System for Video Technology, vol. 6, pp. 191-199, April 1996.

[11] Qin-Fan Zhu, Louis Kerofsky and Marshall B. Garrison, "Low-Delay, Low-Complexity Rate Reduction and Continuous Presence for Multipoint Videoconferencing ," IEEE Transactions on Circuits and Systems for Video Technology, vol. 9, no. 4, pp.666-676, June, 1999.

[12] Jeongnam Youn, Ming-Ting Sun and Chia-Wen Lin, "Motion estimation for high performance transcoding," IEEE Transactions on Consumer Electronics, vol. 44, pp. 649-658, August 1998.

[13] Jenq-Neng Hwang, Tzong-Der Wu and Chia-Wen Lin, "Dynamic frame-skipping in video transcoding," 1998 IEEE Second Workshop on Multimedia Signal Processing, pp. 616 – 621, 1998.

[14] G. Keeman, R. Hellinghuizen, F. Hoeksema and G. Heideman, "Transcoding of MPEG-2 bitstreams," Signal Processing: Image Communication, vol. 8, pp. 481-500, September 1996.

[15] H.T Chen, P.C. Wu, Y.K Lai and L.G. Chen, 'A multimedia video conference system: using region base hybrid coding', IEEE Trans. on Consumer Electronics, Vol. 42, No.3, 1996, pp.781 -786.

[16] Y.H. Chan and W.C. Siu, 'General approach for the realization of DCT/IDCT using convolutions', Signal Processing, Vol. 37, No.3, 1994, pp.357-364.

[17] N.F. Law and W.C. Siu, 'Successive Structural Analysis Using Wavelet Transform for Blocking Artifacts Suppression', revised version submitted to Signal Processing, Jan 2001.

[18] N.F. Law and W.C. Siu, 'Progressive Image Coding based on Visually Important Features', Vol. II, ICIP, 1999, Japan, pp.362-366.

[19] Youhong Lu and Joel M. Morris, "Gabor Expansion for Adaptive Echo Cancellation", IEEE Signal Processing magazine, Vol. 16, No.2, March 1999,pp.68-pp.72)

[20] ITU-T Recommendation G.723.1, "Dual rate speech coderfor multimedia communications transmitting at 5.3 and 6.3 kbit/s," March 1996

[21] A.S. Spanias, "Speech coding: A tutorial review," Proc. IEEE, vol 82, pp.1541-1582, October 1994.

[22] F.A. Westall and S.F.A.Ip, "Digital Signal Processing in Telecommunication", pp.280-283, 308-319

[23] Den-Yuen Hsiau and Ja-Ling Wu, "Real-time PC-based software implementation of H.261 video code," IEEE Transactions on Consumer Electronics, vol. 43, no. 4, pp. 1234-1244, November 1997.

[24] Chwan-Hwa Wu and J. David Irwin, "Multimedia and multimedia communication: A Tutorial," IEEE Transactions on Industrial Electronics, vol. 45, no. 1, pp. 4-14, Feb 1998.

[25] M. Reha Civanlar, Glenn L. Cash, Richard V. Kollarits, Baldine-Brunel Paul, Cassandra T. Swain, Barry G. Haskell and David A. Kapilow, "IP-networked multimedia conferencing," IEEE Signal Processing Magazine, pp. 31-43, July 2000.

[26] P. Lago and G. Canal, "A video-conferencing distributed service tailored for education," IEEE International Conference on Multimedia Computing and Systems, vol. 2, pp. 1038-1042, 1999.

[27] Lynn Conway and Charles J. Cohen, "Video mirroring and iconic gestures: enhancing basic videophomes to provide visual coaching and visual control," IEEE Transactions on Consumer Electronics, vol. 44, no. 2, pp. 388-397, May 1998.

[28] ITU-T Study Group XV, Recommendation H.231, "Multipoint control units for audiovisual systems using digital channels up to 2 Mb/s," May 1992.

[29] ITU-T Study Group XV, Recommendation H.243, "Procedures for establishing communication between three or more audiovisual terminals using digital channels up to 2 Mb/s," May 1992.

[30] Y. Nakajima, H. Hori and T. Kanoh, "Rate conversion of MPEG coded video by re-quantization process," in IEEE International Conference on Image Processing, ICIP95, vol. 3, pp. 408-411, October 1995, Washington, DC.

[31] P. Assuncao and M. Ghanbari, "Post-processing of MPEG2 coded video for transmission at lower bit rates," in IEEE International Conference on Acoustic, Speech, and Signal Processing, ICASSP96, vol. 4, pp. 1998-2001, May 1996, Atlanta, GA.

[32] Y. L. Chan and W. C. Siu, "New adaptive pixel decimation for block motion vector estimation," IEEE Transactions on Circuits and Systems for Video Technology, vol.6, no.1, pp.113 -118, February. 1996.

[33] Y. L. Chan and W. C. Siu, "Edge oriented block motion estimation for video coding," IEE Proceedings: Vision, Image and Signal Processing, vol. 144, no. 3, pp. 136-144, June 1997.

[34] Y. L. Chan and W. C. Siu, "On block motion estimation using a novel search strategy for an improved adaptive pixel decimation," Journal of Visual Communication and Image Representation, vol. 9, no. 2, pp. 139-154, June 1998.

[35] Jo Yew Tham, Surendra Ranganath, Maitreya Ranganath, and Ashraf Ali Kassim, "A novel unrestricted center-biased diamond search algorithm for block motion estimation," IEEE Transactions on Circuits and Systems for Video Technology, vol. 8, pp. 369-377, August 1998.

[36] F.-H Cheng and S.-N. Sun, "New fast efficient two-step search algorithm for block motion estimation," IEEE Transactions on Circuits and Systems for Video Technology, vol. 9, no. 7, pp.977-983, October, 1999.

[37] M.Yong, Q.-F.Zhu, and V.Eyuboglu, "VBR transport of CBR encoded video over ATM networks," in Proc. 6th Int. Workshop Packet Video, Portland, OR, Sept. 1994, pp.D18.1-D18.4.

[38] ISO/IEC 11172-1, "Information Technology -- Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1,5 Mbit/s -- Part 1: System," 1993

[39] E. Weinstein, M. Feder, and A. V. Oppenheim, " Multi-channel signal separation by decorrelation", *IEEE Trans. on Speech and Audio Processing,* Vol. 1, No. 4, pp. 405-413, Oct. 1993.

[40] Masato Abe *et al.*, "Estimation of the Waveform of a Sound Source by Using an Iterative Technique with Many Sensor", *IEEE Trans. on Speech and Audio Processing,* Vol. 6, No. 1, pp.24-35, Jan. 1998.

[41] D. Yellin and E. Weinstein, " Criteria for multichannel signal separation", *IEEE Trans. on Signal Processing,* Vol. 42, No. 8, pp. 2158-2167, Aug. 1994.

[42] C. L. Nikias and A. P. Petropulu, "Higher-order spectra analysis: a nonlinear signal processing framework", Englewood Cliffs, NJ., Prentice Hall, 1993.

[43] ISO/IEC International Standard 11172-3 "Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbits/s – Part 3:Audio", Switzerland, Aug 1993.

[44] Davis Pan, "A Tutorial on MPEG/Audio Compression", *IEEE Multimedia*, vol.2, no.2, pp. 60-74, Summer 1995.

[45] S.Shlien, "Guide to MPEG-1 Audio Standard", *IEEE Transactions on Broadcasting*, vol.40, no.4, pp.206-218, 1994.

[46] Peter Noll, "MPEG digital audio coding", *IEEE Signal Processing Magazine*, pp.59-81, September 1997.

[47] Do-Hui Teh, Soo-Ngee Koh and Ah-Peng Tan, "Efficient bit allocation algorithm for ISO MPEG audio encoder", electronic letters, vol.34, no.8, pp.721-722, April 1998.

[48] Chia-Wen Lin, Te-Jen Liou, and Yung-Chang Chen, "Dynamic bit rate control in multipoint video transcoding," IEEE International Symposium on Circuits and Systems, vol. 2, pp. 17-20, May 28-31, 2000.

[49] L. Chiariglione, "The development of an integrated audiovisual coding standard: MPEG," Proceedings of IEEE, vol. 83, pp. 151-157, February 1995.

[50] H.J. Stuttgen, "Network evolution and multimedia communication," IEEE Multimedia, vol. 2, pp. 42-59, Fall 1995.

[51] S.H. Kwok, W.C. Siu and A.G. Constantinides, "Adaptive temporal decimation algorithm with dynamic time window," IEEE Trans. on Circuits and Syst. for Video Tech., vol.8, pp. 104-111, February 1998.

[52] S.H. Kwok, W.C. Siu and A.G. Constantinides, "A scaleable and adaptive temporal segmentation algorithm for video coding," Graphical Models and Image Proc., vol.59, pp. 128-138, May 1997

[53] Y.H. Chan and W.C. Siu, 'General approach for the realization of DCT/IDCT using convolutions', Signal Processing, Vol. 37, No.3, 1994, pp.357-364.

[54] N.F. Law and W.C. Siu, 'Successive Structural Analysis Using Wavelet Transform for Blocking Artifacts Suppression', revised version submitted to Signal Processing, Jan 2001.

[55] N.F. Law and W.C. Siu, 'Progressive Image Coding based on Visually Important Features', Vol. II, ICIP, 1999, Japan, pp.362-366.

[56] Kai-Tat Fung, Yui-Lam Chan and Wan-Chi Siu, 'Low-Complexity and High Quality Frame-Skipping Transcoder', paper accepted, to be published on proceedings, ISCAS'2001.

[57] M.J. Swain and D.H. Ballard, 'Color Indexing', International Journal of Computer Vision, Vol. 7, No.1, 1991, pp.11-32.

[58] ISO/IEC International Standard 11172-3 "Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbits/s – Part 3:Audio", Switzerland, Aug 1993.

[59] Davis Pan, "A Tutorial on MPEG/Audio Compression", *IEEE Multimedia*, vol.2, no.2, pp. 60-74, Summer 1995.

[60] S.Shlien, "Guide to MPEG-1 Audio Standard", *IEEE Transactions on Broadcasting*, vol.40, no.4, pp.206-218, 1994.

[61] Peter Noll, "MPEG digital audio coding", *IEEE Signal Processing Magazine*, pp.59-81, September 1997.

[62] Do-Hui Teh, Soo-Ngee Koh and Ah-Peng Tan, "Efficient bit allocation algorithm for ISO MPEG audio encoder", electronic letters, vol.34, no.8, pp.721-722, April 1998.