Resource allocation for real-time video coding system

Liu, Yang

2008

Liu, Y. (2008). Resource allocation for real-time video coding system. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/41401>

<https://doi.org/10.32657/10356/41401>

Resource Allocation for Real-Time Video Coding System

LIU Yang

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in fulfillment of the requirement for the degree of
Doctor of Philosophy

2008

To my Parents:

Liu Rong Chao and Yang Ya Min.

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my supervisor, Prof. Soh Yeng Chai, and my co-supervisor, Dr. Li Zheng Guo, for their valuable advice, guidance and inspirational enthusiasm for my research. Their unique insights and perspectives have enriched my research experience. I highly appreciate their kind patience and instant response to my queries even in their busy schedule. Their enthusiasm for the research is a constant source of encouragement throughout my entire candidature.

I wish to extend my appreciation to my friends in the Nanyang Technological University, for their valuable support and help during those long hours spent in the laboratory, and for their kind friendship in the daily life.

Finally, I take this opportunity to thank my mother and father, for the tremendous love, enormous support and constant encouragement they have given me since the first day of my birth.

Contents

Acknowledgements	iv
Summary	viii
List of Figures	x
List of Tables	xvii
1 Introduction	1
1.1 Background	1
1.2 Motivations	3
1.3 Major Contributions	6
1.4 Thesis Outline	8
2 Overview of H.264/AVC and its Scalable Extension	11
2.1 Hybrid Video Coding	11
2.2 H.264/AVC	14
2.3 Scalable Extension of H.264/AVC	23
2.3.1 MCTF	24
2.3.2 Hierarchical B Frame	28

2.3.3	SNR Scalability	28
2.3.4	Spatial Scalability	29
2.3.5	Temporal Scalability	32
2.3.6	Combined Scalability	33
3	Bit Allocation of Rate Control for H.264/AVC	35
3.1	Introduction	35
3.2	Rate Distortion Optimization in H.264/AVC	41
3.3	Improved MAD Prediction	43
3.4	Sum Bit R-Q model	49
3.5	Optimized QP Adjustment at MB level	55
3.6	Proposed Rate Control Algorithm	59
3.7	Experimental Results	71
3.8	Conclusions	81
4	Region-Of-Interest based Resource Allocation for Conversational Video Communication of H.264/AVC	83
4.1	Introduction	83
4.2	Fast ROI Detection	87
4.2.1	Dilemma in the Detection of ROI	87
4.2.2	Simple ROI Detection Method for Conversational Video Communication	88
4.3	ROI Based Bit Allocation for H.264/AVC	95
4.4	ROI Based Computing Power Allocation	96
4.4.1	Encoder complexity	97
4.4.2	Decoder complexity	108
4.5	Experimental Results	111

Contents

vii

4.6	Conclusions	117
5	Bit Allocation of Rate Control for H.264/AVC Scalable Extension	119
5.1	Introduction	119
5.2	Cross SNR Layer ME/MC Scheme	122
5.3	Bit Allocation of Rate Control for SNR and Spatial Enhancement Layer	129
5.4	Bit Allocation of Rate Control for Hierarchical B Frame	131
5.5	Bit Allocation of Rate Control for Combined Enhancement Layer .	134
5.6	Experimental Results	135
5.7	Conclusions	147
6	Conclusions and Future Research Directions	149
6.1	Conclusions	149
6.2	Future Research Directions	151
	Author's Publications	155
	Bibliography	157

Summary

A real-time communication requires a video coding strategy with high compression ratio to achieve the best video quality at a given bitrate, small buffer for low-delay interactive communication, low computing complexity for easy implementation with personal video application terminals (e.g. handphone and PDA), and high scalability to provide an embedded bitstream with adaptive capability for various video quality, frame rate and spatial resolution. This thesis investigates the existing video coding standards, especially the newest H.264/AVC and its scalable extension. To produce the best visual quality of the reconstructed video with a given coding resource constraint and help meet the requirements of real-time communication, three aspects of resource allocation are considered in this thesis.

The first aspect is bit allocation to regulate the generated bitstream at a given bandwidth and to prevent the buffer from overflowing and underflowing. This could be realized with an efficient rate control scheme. After analyzing the problems (inaccurate mean absolute difference (MAD) prediction and inaccurate texture bits prediction) of the existing rate control scheme for H.264/AVC, a rate control scheme with adaptive MAD prediction and sum bits R-Q model is proposed for low delay application of H.264/AVC standard.

The second aspect of this thesis is region-of-interest (ROI) based resource allocation. Most of the video encoders give equal importance to every part of the video regardless of its relative importance to the human visual system (HVS). However, in many video applications, clients might pay more attention to the region of their interest, so more coding resources, including bits and computing power, should be allocated to the parts that will capture more visual attention. An ROI based resource allocation schemes for H.264/AVC standard is developed in this thesis. With the direct frame difference and skin-tone information, the ROI could be detected quickly. Then the ROI based bits allocation is achieved by adjusting the quantization parameter with rate control process, and the computing power allocation is realized by adaptively adjusting other coding parameters, such as the candidates for mode decision, the number of referencing frames, the accuracy of motion vectors and the search range of motion estimation, according to the relative importance of each MB. Moreover, to achieve a decoding-friendly encoder, the decoding complexity could also be optimized at the encoder side by utilizing an ROI based rate-distortion-complexity (R-D-C) cost function. In this way, the encoding resources could be allocated properly and the overall subjective visual quality could also be improved.

The third aspect is bit allocation of rate control for scalable video coding. The bit allocation for fixed bitrate coding and buffer control are also required at the scalable enhancement layers. This thesis presents a rate control scheme for H.264/AVC scalable extension. With our algorithm, the rate control for all the SNR (quality), spatial, and temporal enhancement layers could be achieved. The encoder could encode all the scalable layers into bitstreams with fixed bitrates, and the buffer is well controlled by our scheme to prevent its overflowing and underflowing.

Summary

x

In comparison to the existing video coding schemes, the video coder with our proposed resource allocation strategy achieves better subjective and objective visual quality, the coding complexity of the video coding system is also significantly reduced, and the buffer is well controlled for low delay real-time application.

List of Figures

2.1	From video sequence to block unit.	13
2.2	Backward motion prediction for P frames (up) and bi-directional motion prediction for B frames (low).	13
2.3	The encoder side of hybrid video coding scheme.	14
2.4	The decoder side of hybrid video coding scheme.	14
2.5	Compare the coding efficiency of H.264/AVC and other standards. “H.263 HLP” is the H.263 High Latency Profile, “ASP” is the MPEG-4 Advanced Simple Profile, “H.26L” is Test Model Long Term which is finalized as H.264.	23
2.6	The MCTF based temporal decomposition for one GOP.	27
2.7	An example of hierarchical B frame prediction structure with four temporal levels.	28
2.8	The flow chart for spatial scalable extension.	30
2.9	The up-sampling method of MB partition.	31
2.10	An example of temporal scalability.	32
2.11	An example of combined scalability.	34

List of Figures

xii

3.1	The flowchart of rate control for previous standards (MPEG-1, MPEG-2, H.261, H.263).	41
3.2	The flowchart of rate control for H.264/AVC.	42
3.3	The comparison of actual MAD and direct MAD at frame level. Analysis of the P frames from CREW@QCIF-15Hz and FOOTBALL@QCIF-15Hz sequences, IPPP structure with QP=40.	44
3.4	The comparison of actual MAD, linear predicted MAD and proposed adaptively predicted MAD at frame level. Analysis of the P frames from CREW@QCIF-15Hz and FOOTBALL@QCIF-15Hz sequences, respectively. IPPP structure with QP=40.	46
3.5	The comparison of actual MAD, linear predicted MAD and proposed adaptively predicted MAD at MB level. Randomly analyzed the 67th frame from HARBOUR@QCIF-15Hz, 90th frame from FOREMAN@QCIF-15Hz, 35th frame from FOOTBALL@QCIF-15Hz and 5th frame from CREW@QCIF-15Hz sequences, with QP=40.	48
3.6	The variation of the amount of texture bits and non-texture bits. Analysis of HARBOUR@CIF-15Hz, FOREMAN@QCIF-15Hz, FOOTBALL@QCIF-15Hz and CREW@QCIF-15Hz sequences, with QP=40.	50
3.7	The relationship between sum bit and MAD with fixed QP=40. Analysis of FOOTBALL@QCIF-15Hz and CREW@QCIF-15Hz sequences.	51
3.8	The linear relationship among TextureBit, Non-TextureBit, SumBit and QP. Randomly chosen 1st and 2nd P frames from FOOTBALL@QCIF-15Hz sequence, QP=20:2:40 (QP values are set from 20 to 40, with increase step of 2).	54
3.9	The relation of QP and QS in H.264/AVC.	57

3.10 The linear relationship between PSNR and QP. Randomly chosen 1st, 2nd, 3rd and 4th P frames from FOOTBALL@CIF-15Hz sequence, QP=20:2:40.	58
3.11 The block diagram of our rate control scheme at frame level.	60
3.12 The block diagram of our rate control scheme at MB level.	65
3.13 The comparison of target frame bits mismatch between our scheme and JVT-G012 at MB level rate control. Analysis of FOREMAN@QCIF-15Hz and HARBOUR@QCIF-15Hz sequences, with sequence initial QP=40.	72
3.14 The comparison of PSNR between our scheme and JVT-G012 at MB level rate control. Analysis of FOREMAN@QCIF-15Hz and HARBOUR@QCIF-15Hz sequences, with sequence initial QP=40.	73
3.15 The comparison of buffer status between different coding scheme at MB level rate control. Analysis of CITY@QCIF-15Hz, FOREMAN@QCIF-15Hz, BUS@QCIF-15Hz and MOBILE@QCIF-15Hz sequences, with sequence initial QP=40.	74
3.16 The comparison of reconstructed video quality, the 77th, 78th, 79th and 80th frames of FOREMAN. (a) reconstructed with our scheme, (b) reconstructed with JVT-G012.	76
3.17 The comparison of reconstructed video quality, the 62th, 63th, 64th and 65th frames of FOOTBALL. (a) reconstructed with our scheme, (b) reconstructed with JVT-G012.	76
3.18 The comparison of reconstructed video quality. (a) reconstructed without σ effect, (b) reconstructed with σ effect, (c) details of (a), (d) details of (b).	77

4.1	Example 1 of ROI detection: (a) the 5th original frame of the FOREMAN sequence at QCIF spatial resolution; (b) temporal difference at pixel level; (c) the detected skin-tone areas; (d) the dilated skin-tone area; (e) I_{ROI} at MB level; (f) initial detected ROI mask; (g) final detected ROI mask after mean filtering.	91
4.2	Example 2 of ROI detection: (a) the 22th original frame of the MISS sequence at QCIF spatial resolution; (b) temporal difference at pixel level; (c) the detected skin-tone areas; (d) the dilated skin-tone area; (e) I_{ROI} at MB level; (f) initial detected ROI mask; (g) final detected ROI mask after mean filtering.	92
4.3	Example 3 of ROI detection: (a) the 73th original frame of the M&D sequence at QCIF spatial resolution; (b) temporal difference at pixel level; (c) the detected skin-tone areas; (d) the dilated skin-tone area; (e) I_{ROI} at MB level; (f) initial detected ROI mask; (g) final detected ROI mask after mean filtering.	93
4.4	The boundary ROI mask of one frame.	94
4.5	The MB partitions of H.264/AVC.	98
4.6	The intra prediction methods for 4×4 partition of H.264/AVC . . .	99
4.7	The intra prediction methods for 16×16 partition of H.264/AVC .	100
4.8	The original testing frames are shown in the left column and the final detected ROI mask and the MB partition modes for the same frames are in the right column. Random analysis of (a) the 8th frame of FOREMAN@QCIF 15Hz, (b) the 17th frame of MISS@QCIF 30Hz, (c) the 16th frame of M&D@QCIF 30Hz.	102
4.9	The percentage of the partition subsets for the MBs within the same range of μ value. Analysis of FOREMAN@QCIF 15Hz with QP=40.	103

List of Figures

xv

4.10	The method of multiple reference frames for H.264/AVC.	103
4.11	The method of interpolation for H.264/AVC.	105
4.12	The relation of the percentage of motion vectors belongs to large search range and the μ value. Analysis of FOREMAN@QCIF 15Hz with QP=40.	108
4.13	Example 1 of comparison of the subjective visual quality. Analysis of the 9th frame of FOREMAN@QCIF 15Hz, (a) original frame, (b) detected ROI mask, (c) the reconstructed frame without ROI concern with PSNR = 27.30 dB, (d) the reconstructed frame with ROI concern with PSNR = 26.72 dB, (e) details of the most sensitive region of (c), (f) details of the most sensitive region of (d).	114
4.14	Example 2 of comparison of the subjective visual quality. Analysis of the 30th frame of MISS@QCIF 30Hz, (a) original frame, (b) detected ROI mask, (c) the reconstructed frame without ROI concern with PSNR = 34.04 dB, (d) the reconstructed frame with ROI concern with PSNR = 34.19 dB, (e) details of the most sensitive region of (c), (f) details of the most sensitive region of (d).	115
4.15	Example 3 of comparison of the subjective visual quality. Analysis of the 64th frame of M&D@QCIF 30Hz, (a) original frame, (b) detected ROI mask, (c) the reconstructed frame without ROI concern with PSNR = 27.26 dB, (d) the reconstructed frame with ROI concern with PSNR = 26.61 dB, (e) details of the most sensitive region of (c), (f) details of the most sensitive region of (d).	116
5.1	Cross SNR layer ME/MC scheme.	126
5.2	An example of combined enhancement layer.	134

List of Figures

xvi

5.3	Buffer status with rate control for SNR and spatial scalable video coding of CREW.	142
5.4	Buffer status with rate control for SNR and spatial scalable video coding of FOOTBALL.	143
5.5	Buffer status with rate control for Hierarchical B frame of BUS and CITY.	144
5.6	Buffer status with rate control for combined scalable video coding of CITY.	145
5.7	Buffer status with rate control for combined scalable video coding of FOREMAN.	146
6.1	An example of multiple base layers.	152

List of Tables

2.1	Comparison between Video Coding Standards.	22
3.1	The Percentage of Temporal Model in Use and The Prediction Error of Different Models at Frame Level. IPPP structure with QP=40. .	47
3.2	Comparison of R-Q Model Accuracy. IPPP structure with initial QP=40.	52
3.3	The parameters used in the frame level rate control.	70
3.4	The parameters used in the MB level rate control.	70
3.5	The Comparison of the Target Frame Bits Mismatch.	71
3.6	The Comparison of Final Rate Control Result.	75
4.1	The Calculation Complexity at Different Sub-pixel Position.	107
4.2	Comparison of Our Scheme and JM9.8 with Coding Complexity Optimization.	112
4.3	Comparison of JM9.8 and Our Scheme without Coding Complexity Optimization.	117
5.1	The Setting of SNR and Spatial Scalable Coding.	135
5.2	The Setting of Hierarchical B Frame.	136

List of Tables

xviii

5.3	The Setting of Combined Scalable Coding.	136
5.4	Compare Our Rate Control Scheme with FGS Scheme for SNR and Spatial Scalable Coding.	137
5.5	Compare Our Rate Control Scheme with FGS Scheme for Hierarchical B Frame Coding.	138
5.6	Compare Our Rate Control Scheme with FGS Scheme for Combined Scalable Coding.	138
5.7	Compare Our Rate Control Scheme with Fixed QP Scheme for CGS and Spatial Scalable Coding.	139
5.8	Compare Our Rate Control Scheme with Fixed QP Scheme for Hierarchical B Frames Coding.	140
5.9	Compare Our Rate Control Scheme with Fixed QP Scheme for Combined Scalable Coding.	140

Chapter 1

Introduction

1.1 Background

The demands for multimedia services are rapidly increasing and the expectation of the quality for these services is becoming higher. To attain the highest possible quality and to be free from noise and waveform distortion during transmission and storage, analog signals (such as speech, audio, images, and video) should be sampled and digitized to digital data for transmission/recording and reconstructed at the receiving ends. However, the volume of these directly digitized data is usually huge. Although the technology is continuously pushing up the bandwidth capacity and reducing the transmission or storage cost, the channel bandwidths and storage capacities are still limited and relatively expensive in comparison with the volume of these raw digital data. To make all the digital services feasible and cost effective, data compression is desired. All the digital signals carry redundant information and could be compressed by removing these redundancy for efficient use of bandwidth at the lowest possible cost.

In the area of data compression, video coding have attracted a lot of attention due to many challenging research topics and immediate or potential applications [1],

such as video conferencing, videophone, video on demand, high definition television (HDTV), interactive TV, telemedicine, etc. Video compression is the basic technology for these video applications, and it is achieved by removing the redundancy from the original video sequence. There are three kinds of redundancy within the digital video data: temporal redundancy, spatial redundancy and statistic redundancy. The temporal redundancy could be explored by either the traditional motion estimation and motion compensation technique of MPEG standards [2] or the recently proposed motion compensated temporal filtering (MCTF) [3] [4]. The spatial redundancy could be reduced by transform coding, such as discrete cosine transform (DCT) [5] and discrete wavelet transform (DWT) [6]. The statistic redundancy could be removed by entropy coding methods, such as Huffman coding [7], arithmetic coding [8] and run-level coding (RLC) [9].

With the development of video coding technology, the coding efficiency is improved significantly. The H.264/AVC is the latest international video coding standard which could achieve a significant improvement in rate-distortion efficiency. The H.264/AVC could provide half bit-rate savings when compared with MPEG-2 [10] which is the most common standard used for video storage and transmission, and the compression gain of H.264/AVC over H.263 [11] is in the range of 25% to 50% due to different type of applications. However, such improvement is achieved at the expense of extremely high computing complexity caused by the newly adopted technologies in H.264/AVC.

Besides the coding efficiency, to satisfy the increasing demands of video streaming over computer networks, the transmission of video over heterogeneous networks should be more reliable, which requires efficient coding, as well as scalability to different client capabilities, system resources, and network conditions [12]. For example, clients may have different display resolutions, systems may have different

caching or intermediate storage resources, and networks may have various bandwidths, loss rates, and best-effort or quality of service (QoS) capabilities. Scalable video coding has been proposed to increase the adaptability of video coding system for various networks and client conditions. There are many applications that require scalable video coding: Internet video, wireless LAN video, mobile wireless video for conversational, video on demand (VOD), live broadcasting purposes, end-to-end Internet/wireless video delivery, multi-channel content production and distribution, storage applications, multi-point surveillance systems, and so on. Highly scalable coding is intended to accommodate decoding for various video applications with diverse needs in resolution, video quality and frame rate. The scalable functions provided in the H.264/AVC scalable extension include SNR (quality) scalability, spatial scalability, temporal scalability and combined scalability.

1.2 Motivations

With the fast development of video communication technology, real-time multimedia services, such as live video broadcasting, video telephony and video conferencing, have become important components of daily communication [13]. This kind of real-time communication requires a video coding strategy that can provide feasible solutions to the following problems:

- (a) How to encode the video sequences into bitstreams with given bitrates?
- (b) How to control the buffer to prevent it from overflowing and underflowing?
- (c) How to improve the subjective visual quality with concern of the human visual system?
- (d) How to reduce the computing complexity for easy implementation of the video

coding system with personal video application terminals (e.g. handphone and PDA)?

- (e) How to solve the problems stated in (a) and (b) for scalable video coding?

The motivations of the research in this thesis are to address these problems from the following three aspects:

(1) Bit allocation of Rate Control for H.264/AVC:

The motivation of this part is to solve the problems (a) and (b) stated above.

Bit allocation of rate control has played an important role in video services, especially for real-time communication. In many video applications, the compressed video signal is to be delivered through constant bitrate channels, but the information within each frame is inherently compressed into variable bits. Due to the complex video encoding process, it seems not feasible to adjust only the encoding parameters to allocate extremely accurate fixed bits to every encoded frame. Therefore, one buffer should be added between the video encoder and the transmission channel to regulate the bitstream before transmission. With a proper rate control strategy, the encoding parameters could be adjusted to prevent the buffer from overflowing and underflowing. This is important because buffer overflowing and underflowing will cause frame skipping and wasting of channel resource, respectively.

Every video coding standard has its own rate control technique, such as TM5 [14] for MPEG-2 [10], TMN8 [15] and TMN12 [16] for H.263 [11] [17], VM8 [18] [19] for MPEG-4 Visual part 2 [20], JVT-G012 [21] for H.264/AVC [22] and so on. The rate control for H.264/AVC is more difficult than that for MPEG-4 Visual part 2. This is because the quantization parameter (QP) is used for both rate control and rate distortion optimization (RDO) [23]. That

is to say, the statistic information of residual data (mean absolute difference (MAD)) needed by the rate control algorithm is available only after the RDO has used a QP value to generate it. To solve this QP-MAD dilemma, Li et al. [21] suggested an estimate for MAD based upon the available information of the previous frames. It is a temporal linear model for MAD prediction. The temporal linear model assumes that the statistic information varies gradually from frame to frame. However, if a scene change or high motion occurs, information collected from the previous frames is no longer useful and the temporal linear model fails to predict an accurate MAD. In that case, the rate control result degrades significantly. Meanwhile, in H.264/AVC, the amount of non-texture bits fluctuates with unpredictable property and a higher percentage is required than that of the previous standards (MPEG-1, MPEG-2, H.261, H.263), especially at low bitrates [24]. So the amount of non-texture bits of the current basic unit cannot be simply predicted from the recent encoded basic units, and the error from the simple prediction of non-texture bits will impair the final rate control result. Therefore, we should investigate a more robust bit allocation strategy with better rate control result to overcome the inaccuracies in MAD prediction and non-texture bits estimation.

(2) Region-Of-Interest (ROI) based resource allocation for H.264/AVC:

The main objective of this part is to solve the above stated problems (c) and (d).

Most of the available video encoders allocate bits [14–16,18,19,21] or computing power [25–32] with equal importance to every part of the video regardless of its relative importance to the human visual system (HVS). However, in many video applications, clients might pay more attention to the region of their interest, so more coding resources including bits and computing power

should be allocated to parts that will attract more visual attention. With the ROI based coding resource allocation, the subjective visual quality could be improved with concern of the human visual system, and the computing complexity could be reduced for easy implementation.

(3) Bit allocation of rate control for H.264/AVC scalable enhancement layers:

The main focus in this part is to solve problem (e) mentioned earlier.

Besides H.264/AVC, bit allocation of rate control is also desired for all the SNR, spatial, and temporal enhancement layers. In the current JSVM reference software [33], there is no rate control mechanism for each scalable layer. We could extend our bit allocation strategy for H.264/AVC rate control to the scalable enhancement layers, by taking into account the compression method used for spatial enhancement, temporal enhancement, as well as SNR enhancement layers.

1.3 Major Contributions

The major contributions of this thesis are as follows:

(i) Proposal of a bit allocation scheme for H.264/AVC rate control:

A switched mean absolute difference (MAD) prediction scheme is introduced to enhance the traditional temporal MAD prediction model, which is not suitable to predict abrupt MAD fluctuation. Moreover, an accurate sum bits R-Q model is also formulated to describe the relationship between the total amount of bits for both texture and non-texture information and quantization parameter (QP), so that the negative effect caused by inaccurate non-texture bits estimation is removed. By exploring the relationship between the PSNR

and QP value, our proposed sum bits R-Q model could further optimize the QP calculation at the MB level. When compared with the bit allocation scheme JVT-G012 [21] which is adopted in the latest JVT H.264/AVC reference model JM9.8, our algorithm could reduce the mismatch between actual bits and target ones by up to 75%. To meet the low delay requirement, the buffer is better controlled to prevent it from overflowing and underflowing. The average luminance PSNR of reconstructed video is increased by up to 1.13dB at low bitrates, and the subjective video quality is also improved.

(ii) **Development of ROI based resource allocation schemes for H.264/AVC:**

Due to the complexity of H.264/AVC, it is very challenging to apply this standard to design a conversational video communication system. This problem is addressed in this thesis by using ROI based bit allocation and computing power allocation schemes. In our system, the ROI is first detected by using the direct frame difference and skin-tone information. Several coding parameters including quantization parameter (QP), candidates for mode decision, number of referencing frames, accuracy of motion vectors and search range of motion estimation are adaptively adjusted at the MB (macro-block) level according to the relative importance of each MB. Subsequently, the encoder could allocate more resources such as bits and computing power to the ROI, and the decoding complexity is also optimized at the encoder side by utilizing an ROI based rate-distortion-complexity (R-D-C) cost function. The encoder is thus simplified and decoding-friendly, and the overall subjective visual quality can also be improved. When compared to the latest H.264/AVC reference model JM9.8, our simplified encoder can be used to achieve much better subjective visual quality with comparable PSNR value. The mismatch between the actual frame bits and the target ones is reduced by up to 80%.

Meanwhile, our proposed encoder could save the motion estimation time by up to 73%, save the overall coding time by up to 63% and reduce the decoding interpolation complexity by up to 48%.

(iii) **Presentation of a bit allocation scheme of rate control for H.264/AVC scalable extension:**

Based on our previous result for H.264/AVC rate control, a switched model is proposed to predict MAD by the available MAD information from the previous frame of the current layer and the current frame of the previous layer. Thus, abrupt MAD fluctuations could be predicted properly in the enhancement layer. Moreover, a bit allocation scheme for hierarchical B frame prediction structure is proposed. With our algorithm, the rate control for all the SNR, spatial, and temporal enhancement layer could be realized, and the encoder could achieve fixed bitrate encoding for each scalable layer. When compared with the FGS based fixed bitrate coding method in JSVM 7.9, the coding efficiency is improved by up to 9.4 dB using our rate control algorithm, because the adjusted QP value could achieve a good balance between the motion and texture bits due to rate distortion optimization (RDO). Compared with the fixed QP encoding method, which iterates the coding process for up to 15 times, our scheme encodes the sequence only once. Moreover, to meet the low delay requirement, the buffer is well controlled by our scheme to prevent it from overflowing and underflowing.

1.4 Thesis Outline

The rest of the thesis is organized as follows:

Chapter 2 first overviews the basic ideas behind the hybrid video coding structure. It then summarizes the new features adopted in the H.264/AVC video coding standard, and compares the differences between H.264/AVC and its previous standards in terms of specific features and coding efficiency. Finally, the H.264/AVC scalable extension is briefly introduced at the end of this chapter.

Chapter 3 proposes a bit allocation scheme of H.264/AVC rate control for low delay video communication. The rate distortion optimization (RDO) employed in H.264/AVC standard is first described. Then, the observation of a relationship between direct MAD and actual MAD motivates the proposal of a direct MAD prediction model. A prediction model switching strategy is also introduced to adaptively choose a prediction model with higher accuracy. Moreover, we propose a new sum bits R-Q model to solve the inaccurate texture bit prediction problem, and the quantization parameters could be next optimized at the MB level by Lagrange theory. With the proposed adaptive MAD prediction method and the sum bits R-Q model, our rate control algorithm is then described in detail. Finally, several experimental results and relevant discussions are provided.

Chapter 4 develops the ROI based resource allocation schemes for H.264/AVC. It first presents our ROI detection scheme for H.264/AVC standard. Based on the rate control scheme for H.264/AVC proposed in Chapter 3, an ROI based bit allocation is implemented. Meanwhile, the computing power could also be properly allocated with the determined ROI information. At the end of this chapter, our simplified system is compared with the JVT H.264/AVC reference software JM9.8 with experimental results and discussions.

Chapter 5 presents a bit allocation scheme for rate control of H.264/AVC scalable extension. Based on the rate control scheme for H.264/AVC proposed in Chapter 3, the bit allocation algorithm for hierarchical B frame, SNR, spatial and

temporal enhancement layers are respectively described in this chapter. The efficiency of our rate control scheme for H.264/AVC scalable extension is supported by experimental results.

Chapter 6 concludes this thesis with concluding summaries and discussions for future research directions.

Chapter 2

Overview of H.264/AVC and its Scalable Extension

2.1 Hybrid Video Coding

Although there exists several other video coding schemes, such as the wavelet based sub-band video coding [34] [35] [36], the hybrid video coding [10] is the most common algorithm with successful applications. In the hybrid video coding scheme, the reconstructed previous frame is used as the reference for motion estimation and motion compensation to remove temporal redundancy. The resulting displaced frame difference, which has much lower energy as compared with the original frame, is then processed with discrete cosine transform (DCT) to remove the spatial redundancy.

The data structure in a video sequence is shown in Figure 2.1. Based on the group of pictures (GOP) structure, the encoder side needs to determine whether the incoming frame is coded as an intraframe (I) or an interframe (P/B). Each I-frame is then divided into blocks, each block of pixels is converted through a DCT unit to DCT coefficients, which are then quantized according to a quantization

table. The DC coefficients of neighboring blocks are arranged in a group and processed by means of differential pulse code modulation (DPCM) to generate differential DC values. The AC coefficients are reordered in a zigzag way and then translated to entropy coding. For B/P frames, as shown in Figure 2.2, motion estimation could be performed on each block, based on the current frame and the reconstructed reference frame. For B frames, both the forward and the backward motion predictions are needed. With the determined motion vectors, predictive motion compensation is used on the reconstructed reference frame to generate the motion compensated frame residual. This residual is then passed through the DCT unit and a quantization unit to generate a set of quantized DCT coefficients. All the coefficients are next coded in the same way as the intraframe. The determined motion vectors are first coded with the DPCM method, and then coded with entropy coding [1].

The hybrid coding algorithm has been employed in the present digital video compression standards, e.g., H.261 [37], H.263 [11] [17], MPEG-1 [38], MPEG-2 [39], MPEG-4 Visual part 2 [40] [41] and H.264/AVC [42] [22]. H.261 [37] is developed by CCITT (Consultative Committee for International Telephone and Telegraph) in 1988-1990, designed for video conferencing and video-telephone applications over ISDN telephone lines, and the targeting bitrate is $p \times 64$ ($1 \leq p \leq 30$) kb/sec. H.263 [11] [17] was proposed in 1995 and designed for low bitrates (less than 64 kb/sec) communication. MPEG-1 [38] was released in 1991 targeting the VHS (Video Home System) quality on a CD-ROM or VCD. Unlike MPEG-1 which is basically a standard for storing and playing video on a single computer at relatively low bitrates, MPEG-2 [39] is a standard proposed in 1994 for digital TV and meets the requirements for HDTV (High Definition Television) and DVD. MPEG-4 Visual part 2 video coding [40] [41] is the first standard to consider

Figure 2.1: From video sequence to block unit.

content-based coding and introduces the concept of video object, each video frame could be segmented into several objects and coded separately. MPEG-4 Visual part 2 supports all functions which are already provided by MPEG-1 and MPEG-2. Although these standards are operating at different bitrates and targeting for different applications, they all follow a similar hybrid video coding framework in terms of the coding algorithms. The encoder side and decoder side of the hybrid

Figure 2.2: Backward motion prediction for P frames (up) and bi-directional motion prediction for B frames (low).

Figure 2.3: The encoder side of hybrid video coding scheme.

Figure 2.4: The decoder side of hybrid video coding scheme.

video coding frame work is respectively described in Fig. 2.3 and Fig. 2.4.

2.2 H.264/AVC

In 1995, after finalizing the H.263 standard, the ITU-T Video Coding Experts Group (VCEG) started the work to develop a new standard for low bitrate visual communication. In 1998, VCEG issued a call for proposal on a project called H.26L,

with the target to double the coding efficiency as compared with the previous ITU-T standards. In 2001, the Joint Video Team (JVT), which includes the experts from VCEG and MPEG (the ISO Motion Picture Experts Group), was formed to realize the H.26L draft into a real International Standard. The official title of this new standard is Advanced Video Coding (AVC) [42] [22], and it is also well known by its previous working name, H.26L and H.264. In fact, there are other two identical titles --- AVC: ITU-T H.264 and ISO MPEG-4 Part 10 [42] [22]. In this thesis, this new standard is simply specified by the name “H.264/AVC”.

Besides providing higher coding efficiency than previous standards (MPEG-1, MPEG-2, MPEG-4 Visual part 2, H.261 and H.263), the H.264/AVC standard is a straightforward video coding strategy designed to achieve a network-friendly video representation, simple syntax specifications, seamless integration of video coding into all current protocols and more error robustness. H.264/AVC could be employed in a wide range of applications for video content including but not limited to the following [42] [22]:

- Direct broadcast (DBS) satellite video services;
- Multimedia services over packet networks (MSPN);
- Real-time conversational (RTC) services (videoconferencing, videophone, etc.);
- Remote video surveillance (RVS);
- Digital terrestrial television broadcasting (DTTB), cable modem, DSL;
- Interactive storage media (ISM) (optical disks, etc.);
- Multimedia mailing (MMM);
- Serial storage media (SSM) (digital VTR, etc.);

- Digital Cinema, Content contribution, content distribution, studio editing, post processing.
- Video Streaming over the internet;
- Cable TV (CATV) on optical networks, copper, etc.;

In practice, a complete encoder and decoder of H.264/AVC are likely to include the functional elements shown in Fig. 2.3 and Fig. 2.4. Most of the basic functional elements (prediction, transform, quantization, entropy encoding) of H.264/AVC are also presented in the previous video coding standards (MPEG-1, MPEG-2, MPEG-4 Visual part 2, H.261, H.263) but the important changes in H.264/AVC occur in the details of each function block. Compared to the previous video coding methods, some features to enable enhanced coding efficiency are newly introduced in H.264/AVC, including but not limited to the following [42] [22]:

- *Multiple reference frame motion estimation and compensation:*

In MPEG-2 and its previous standards, the motion estimation and compensation is carried out between the current frame and the previously reconstructed frame. H.264/AVC extends the enhanced reference frame selection method used in H.263. It enables the encoder to choose the reference frame from multiple reconstructed frames with highest predicting efficiency. This method with multiple reference frames could be used in both single directional and bi-directional motion estimation.

- *Weighted prediction:*

For some abnormal video sequence such as the one with fades, the motion compensation method used in the previous standards does not work well. H.264/AVC allows the motion-compensated residual data to be weighted

and offsetted by the amounts specified by the encoder. This can significantly improve the coding efficiency for the sequence containing fades, and it can also be used for other purposes.

- *Directional spatial prediction for intra coding:*

More flexible intra prediction method is provided in H.264/AVC to explore the correlation between the edges of the previously reconstructed neighboring macro-blocks and the current coding macro-block. This method improves the predicting efficiency when the temporal correlation is not effective for prediction.

- *Decoupling of referencing order from display order:*

In the previous standards, there is a strict dependency between the ordering of motion estimation reference frames and the ordering of frames to be displayed. In H.264/AVC, this restriction is removed by allowing the encoder to choose the ordering of reference frames and displaying frames with a high degree of flexibility, only with the constraint of a total memory capacity limit to ensure decoding ability.

- *In-loop de-blocking filter:*

In the block-based video coding algorithms, the blocking artifacts (due to block-based motion compensation and quantization) significantly adverse the subjective visual quality. Based on a concept from an optional feature of H.263, the de-blocking filter in the H.264/AVC design improves both objective and subjective visual quality of the reconstructed frames. Because the reconstructed frames serve as the reference frames for further motion estimation, the de-blocking filter also increases the temporal prediction efficiency.

- *Quarter-sample-accurate motion estimation and compensation:*

The quarter-sample-accurate motion estimation and compensation was first used in the advanced profile of the MPEG-4 Visual part 2 standard, and most of the previous standards only enable half-sample-accurate motion estimation and compensation. The new design of H.264/AVC employs the quarter-sample-accurate motion vector by reducing the interpolation complexity compared to the advanced profile of the MPEG-4 Visual part 2.

- *Variable partition motion estimation and compensation with small block size:*

The MB partition of H.264/AVC varies from 16x16 to 4x4, it provides more flexible MB partition choice for motion estimation and motion compensation than any previous standard.

- *Motion vectors over frame boundaries:*

The frame boundary extrapolation method, which was first used in H.263, is also employed in H.264/AVC. In the MPEG-2 and its previous video coding standards, only the pixels within the previous reconstructed frames are used as reference for motion estimation and compensation.

- *Decoupling of frame displaying methods from frame referencing capability:*

In H.264/AVC, the bi-directional predicted frames could also be used as references for prediction of other frames, they are forbidden in the previous standards. In this way, H.264/AVC provides the encoder with more flexibility to choose the reference frame with higher prediction efficiency.

- *Improved “skipped” and “direct” motion modes:*

In H.264/AVC, the “skipped” mode of motion estimation could contain motion information to improve the predicting efficiency for the sequences with global motion. Otherwise, the motion information is not included for the

“skipped” area in the previous standards. Meanwhile, the “direct” mode is newly introduced in H.264/AVC for bi-directional motion estimation.

- *Arithmetic entropy coding:*

As arithmetic coding was previously found as an optional feature of H.263, H.264/AVC provides an advanced arithmetic entropy coding method, named CABAC (context-adaptive binary arithmetic coding).

- *Context-adaptive entropy coding:*

In H.264/AVC, two entropy coding methods are employed: CABAC and CAVLC (context-adaptive variable-length coding). Both these entropy coding methods are adaptively context-based and outperforms the entropy coding methods used in previous standards.

- *Small block-size transform:*

In the previous video coding standards, the block transform size is fixed to 8x8. To reduce the “ringing” artifacts, H.264/AVC enables the encoder to transform the spatial signal data based on a 4x4 block. Meanwhile, the smaller block size also corresponds to a smaller macro-block partition for motion estimation and compensation.

- *Hierarchical block transform:*

Although the small 4x4 block size transform is perceptually beneficial in most cases, it is still desired to provide some transforms with longer basis functions for the signal with sufficient correlations. H.264/AVC uses a hierarchical transform to extend the effective block size for low frequency chroma information to 8x8. For intra coding, the length of the luminance transform for low-frequency information could be extended to a 16x16 block size.

- *Short word-length transform:*

H.264/AVC standard only requires 16-bit calculation for transform, while the previous ones usually requires 32-bit processing. In this way, the short word-length induces a simple transform calculation.

- *Exact-match inverse transform:*

Based on an optional feature in the H.263, H.264/AVC is the first standard to achieve exact equality of decoded video content from all decoders. In previous video coding standards, because the exact match to the ideal specified inverse transform is impractically achieved, the DCT transform is generally specified within an error tolerance boundary. This causes a drift problem between the encoder and decoder, so each decoder design would produce slightly different reconstructed video.

- *Redundant frames:*

The H.264/AVC standard allows the encoder to send some regions of the frame multiple times when some important parts of the frame is lost during transmission. In this way, the coding system is more robust to data loss.

- *Data Partitioning:*

Because some data, such as the header and motion information is more important than others in terms of reconstructing the video content, the H.264/AVC standard provides the syntax of each slice to be separated into up to three different partitions for transmission. This method is based on some previous work of MPEG-4 Visual part 2 and an optional part of H.263.

- *SP/SI synchronization/switching frames:*

The H.264/AVC standard introduces some new frame types that allow synchronization function of the decoding process of some decoders with the help

from other decoders, rather than decreasing the coding efficiency of all decoders by sending an I frame.

- *Parameter set structure:*

Because the loss of a few key bits of header information (such as sequence header or frame header information) could induce a disastrous impact on the decoding process of the previous standards, the H.264/AVC provides a parameter set structure to separate the header information and handles them in a more flexible and specialized manner.

- *Unit syntax structure:*

Each syntax structure in H.264/AVC is placed into a logical data packet called a NAL (Network Abstraction Layer) unit. Unlike forcing a specific bitstream interface to the system as in the previous video coding standards, the NAL unit syntax structure allows more flexibility of carrying the video content in a manner appropriate for each specific network.

- *Flexible macro-block ordering (FMO):*

In H.264/AVC, a new ability to partition the frame into regions called slice groups is newly introduced. In this way, each slice could be decoded independently as a subset of a slice group, and this method can significantly enhance robustness to data losses by managing the spatial relationship between the regions that are coded in each slice.

- *Arbitrary slice ordering (ASO):*

Because each slice of a coded frame can be decoded independently on the other slices of the frame, the H.264/AVC standard allows the sending and receiving of the slices of the frame without constrained order. This technique

2.2 H.264/AVC

22

is based on an optional part of H.263, and it could improve the end-to-end delay for real-time applications.

- *Flexible slice size:*

In MPEG-2, the slice structure is fixed and the coding efficiency is reduced by decreasing the effectiveness of prediction and increasing the quantity of header data. These slice sizes in H.264/AVC are highly flexible as earlier standard MPEG-1.

Table 2.1: Comparison between Video Coding Standards.

Feature/Standard	MPEG-1	MPEG-2	MPEG-4 Visual part 2	H.264/AVC
Macro-block size	16x16	16x16 (frame mode) 16x8 (field mode)	16x16	16x16
Block Size	8x8	8x8	16x16, 16x8, 8x8	16x16, 8x16, 16x8, 8x8, 4x8, 8x4, 4x4
Transform	8x8 DCT	8x8 DCT	8x8 DCT/wavelet	4x4, 8x8 Int DCT 4x4, 2x2 Hadamard
Quantization	Scalar quantization with step size of constant increment	Scalar quantization with step size of constant increment	Scalar quantization with step size of constant increment	Scalar quantization with step size increase at the rate of 12.5%
Entropy coding	VLC	VLC	VLC	VLC, CAVLC, CABAC
Motion Estimation & Compensation	Yes	Yes	Yes	Yes, more flexible Up to 16 MVs per MB
Playback & Random Access	Yes	Yes	Yes	Yes
ME accuracy	Integer, 1/2-sample	Integer, 1/2-sample	Integer, 1/2-sample, 1/4-sample	Integer, 1/2-sample, 1/4-sample
Profiles	No	5	8	4
Reference frame	One	One	One	multiple
Bidirectional prediction mode	forward/backward	forward/backward	forward/backward	forward/forward forward/backward backward/backward
Frame Types	I, P, B, D	I, P, B	I, P, B	I, P, B, SP, SI
Error robustness	Synchronization & concealment	Data partitioning, FEC for important packet transmission	Synchronization, Data partitioning, Header extension, Reversible VLCs	Data partitioning, Parameter setting, Flexible macroblock ordering, Redundant slice, Switched slice
Transmission rate	Up to 1.5Mbps	2-15Mbps	64kbps - 2Mbps	64kbps -240Mbps
Compatibility with previous standards	n/a	Yes	Yes	Yes
Encoder complexity	Low	Medium	Medium	High

Figure 2.5: Compare the coding efficiency of H.264/AVC and other standards. “H.263 HLP” is the H.263 High Latency Profile, “ASP” is the MPEG-4 Advanced Simple Profile, “H.26L” is Test Model Long Term which is finalized as H.264.

With the above mentioned features, H.264/AVC outperforms previous standards in terms of coding efficiency and network adaption. One example to compare the coding efficiency of H.264/AVC and other standards is given in [43] and as shown in Fig. 2.5. The most significant features that contribute to the superior rate-distortion performance of H.264/AVC are the flexible motion estimation strategy and the context based arithmetic coding scheme. The comparison of specific coding algorithms among H.264/AVC and other previous video coding standards is also given in Table 2.1 [44].

2.3 Scalable Extension of H.264/AVC

Although the H.264/AVC standard provides high coding efficiency at a fixed bitrate, when video data is transmitted over error-prone and variable bandwidth channels, it is desirable to provide an embedded bitstream with highly adaptive capability for various video quality, frame rate and spatial resolution. A new promising technology providing these features is named scalable video coding

(SVC), which is under development as MPEG-21 part 13. During the 68-th MPEG meeting in March 2004, 15 proposals were received for the SVC request. The MPEG meeting decided to choose the Work Draft (WD) from two most impressive schemes: wavelet based three-dimensional sub-band video coding [34] (proposed by MicroSoft Research Asia), and scalable extension of H.264/AVC [45] (proposed by Fraunhofer Institute for Telecommunications - Herinrich Hertz Institute). After improvements and comparisons for about more than half a year, in January 2005, the MPEG meeting finally selected the scalable extension of H.264/AVC [46] as the Scalable Video Model (SVM) reference software.

In this thesis, our contributions for scalable video coding are based on the scalable extension of H.264/AVC proposed by Fraunhofer Institute for Telecommunications - Herinrich Hertz Institute (HHI). Compared to the H.264/AVC standard, its scalable standard provides SNR, spatial and temporal scalability, and the methods of MCTF (Motion Compensated Temporal Filtering) and “hierarchical B frame” are employed as two options to remove temporal redundancy.

2.3.1 MCTF

Because of a high similarity between the adjacent frames, efficient video coding technique requires an effective removal of temporal redundancy. The traditional video compression method is hybrid coding [10], which employs closed-loop motion compensation in the temporal domain and 2-D discrete cosine transform (DCT) algorithm in the spatial domain, as described in Section 2.1. This method has been taken in the present digital video compression standards, e.g., H.261, H.263, MPEG-1, MPEG-2, MPEG-4 Visual part 2. In this scheme, the reconstructed previous frame is used to predict the current frame before motion compensation. The resulting displaced frame difference, which has much lower energy as compared

with the original frame, is then encoded and transmitted to the decoder side.

There are some disadvantages in this method. First, the reconstructed reference frame has lower quality than the original frame, especially at low bitrate. So using this distorted reconstructed frame to perform ME/MC can lead to decrease in the temporal prediction efficiency. Second, the closed-loop based video coding scheme employs layered coding and bitplane scheme to achieve quality (SNR) scalability, such as fine granular scalability (FGS) method [47] and its improvement versions [48–50]. Due to only the reconstructed frames in the base layer are used as reference for motion prediction, the prediction gain is relatively low at high bitrate. Third, if there is some error occurring during video transmission, the error will propagate (the quality of the subsequent reconstructed frames will keep decreasing) until the next intra(I)-frame is received. So this closed-loop ME/MC scheme is not suitable for the heterogeneous network environments, such as internet.

To overcome such disadvantages, open-loop ME/MC based video coding is proposed and becoming the focus of attention in scalable video coding. There is one more option to remove temporal redundancy for the scalable video coding, which is called motion-compensated temporal filtering (MCTF). MCTF was first introduced by Ohm [4] and later improved by Choi and Woods [3]. By derivation from the lifting scheme [35, 36], MCTF ensures perfect reconstruction of the input even when non-linear operations are used during the lifting process.

At the decomposition (analysis) side of MCTF, the input signal is L^n , where n is the temporal decomposition level ($n = 0$ for the original frames). The temporal high-pass and low-pass subbands at temporal level $n + 1$ are respectively derived by the following equations (Note that for simple expression, the spatial coordinate and the corresponding motion vectors are not specified):

$$H^{n+1}[k] = L^n[2k+1] - P(L^n[2k]) \quad (2.1)$$

$$L^{n+1}[k] = L^n[2k] + U(H^{n+1}[k]) \quad (2.2)$$

where k is the temporal coordinate for each temporal level, $H^{n+1}[k]$ and $L^{n+1}[k]$ are the temporal high-pass and low-pass subbands at temporal level $n+1$, and $P(\bullet)$ and $U(\bullet)$ are prediction and update operations, which are different for Haar or 5/3 transform based temporal decomposition:

$$P_{Haar}(L^n[2k]) = L^n[2k] \quad (2.3)$$

$$P_{5/3}(L^n[2k]) = \frac{1}{2}(L^n[2k] + L^n[2k+2]) \quad (2.4)$$

$$U_{Haar}(H^{n+1}[k]) = \frac{1}{2}H^{n+1}[k] \quad (2.5)$$

$$U_{5/3}(H^{n+1}[k]) = \frac{1}{4}(H^{n+1}[k] + H^{n+1}[k-1]) \quad (2.6)$$

It can be seen from the above equations that Haar transform based temporal decomposition only covers temporal information of one directional neighboring frame, on the other hand 5/3 transform based temporal decomposition covers both directions. The uni-directional ME/MC generates less motion data as compared with bi-directional ME/MC. It could save more bits for the spatial information at a given bitrate. However uni-directional ME/MC explores less temporal correlation, this increases the energy of residual data in temporal high-pass subband. Therefore, to balance this pair of opposite effects and to achieve high coding efficiency, uni-directional and bi-directional MCTF should be adaptively employed at macro-block level for temporal decomposition.

2.3 Scalable Extension of H.264/AVC

27

The MCTF structure of for one GOP with 12 frames is shown in Fig. 2.6, where H^n and L^n are the temporal high-pass and low-pass subbands at temporal level n .

Figure 2.6: The MCTF based temporal decomposition for one GOP.

The reconstruction (synthesis) equations for MCTF are easily derived from the decomposition (analysis) side:

$$L^n[2k] = L^{n+1}[k] - U(H^{n+1}[k]) \quad (2.7)$$

$$L^n[2k+1] = H^{n+1}[k] + P(L^n[2k]) \quad (2.8)$$

Figure 2.7: An example of hierarchical B frame prediction structure with four temporal levels.

2.3.2 Hierarchical B Frame

When the temporal update procedure is removed from 5/3 wavelet based MCTF, a predictive video coding structure is formed with hierarchical B frame. An example of hierarchical B frame prediction structure with four temporal levels is shown in Fig. 2.7. In one GOP, the key frame (e.g. 0-th, 8-th, 16-th display order in Fig. 2.7) [51], is first coded as an intra frame or inter frame with prediction from the key frame of the previous GOP. The remaining frames are hierarchical bi-directionally predicted, the coding order within one GOP is “ I_0/P_0 , B_1 , B_2 , B_2 , B_3 , B_3 , B_3 , B_3 ”. The key frames are only predicted from other key frames, and the non-key frames are predicted from the nearest frames of the lower temporal level. In this way, the hierarchical B frame structure could also provides the temporal scalability as MCTF.

2.3.3 SNR Scalability

For the SNR base layer, H.264/AVC based transform coding is used. The predicted frames contain intra or residual macro-blocks as in the hybrid video coding described in the Section 2.1. The residual macro-blocks are encoded with DCT and

quantization, and the intra macro-blocks are coded using the intra coding modes, the same as H.264/AVC. The intra frames are coded independently of each other as H.264/AVC intra frames. The SNR base layer provides the minimal visual quality with an initial quantization step size.

On top of the SNR base layer, SNR enhancement layers are coded using an embedded quantization approach [46]. The quantization error between the SNR base layer and the original predicted frames is re-quantized with a smaller quantization step size, and further coded with the same remaining procedures as for the base layer. This enhancement layer together with its corresponding base layer could be combined as the base layer for the next enhancement layer encoding, and this FGS refinement could be applied repeatedly with additional SNR enhancement layers. In this way, multiple SNR enhancement layers could be achieved by repeatedly decreasing the quantization step size and applying DCT transform with a corresponding entropy coding. The SNR enhancement layers provide progressive refinement to the SNR base layer by the bit plane coding method. The progressive refinement can be truncated at any arbitrary point, so that the quality of the SNR base layer can be improved in a fine granular way, similar to the fine granular scalable work proposed for MPEG-4 Visual part 2 [47].

2.3.4 Spatial Scalability

For spatial enhancement layer, the inter-layer prediction is introduced in the H.264/AVC scalable extension to remove the inter-layer correlation with the base layer, as shown in Fig. 2.8 [52].

There are three parts of the inter-layer prediction:

Figure 2.8: The flow chart for spatial scalable extension.

- (i) *Inter-layer motion prediction*: To predict the motion vectors from the available motion information of the lower resolution spatial layer, two more macro-block modes are utilized besides the MB partitions used in H.264/AVC: “BASE_LAYER” and “QPEL_REF” [46]. For these two modes, the macro-block partition is obtained by up-sampling the partition of the corresponding 8x8 block of the lower resolution layer, as shown in Fig. 2.9 [46]. The reference frame for the macro-block in the enhancement layer is the same as the corresponding sub-macro-block partition of the base layer, and the associated motion vectors are scaled by a factor of 2. While for MB with “BASE_LAYER” mode, no additional motion information is coded or transmitted, for the “QPEL_REF” mode, a quarter-sample motion vector refinement is further transmitted for each motion vector. Meanwhile, the up-sampled motion vector from the lower spatial resolution layer could also be used as the motion vector predictor for the MB of the spatial enhancement layer. A flag would be transmitted with each motion vector difference to indicate whether the motion vector predictor is estimated from the spatial neighboring MBs or from the corresponding scaled base layer motion vector.

Figure 2.9: The up-sampling method of MB partition.

- (ii) *Inter-layer intra texture prediction:* In order to enable the inter-layer prediction of low-pass signals, an additional intra macro-block mode is introduced [46]. In this mode, the predicted intra texture information is generated by up-sampling the reconstructed intra texture information from the lower resolution layer using the 6-tap filter which is defined in H.264/AVC for the purpose of half-sample interpolation. Note that, this intra texture prediction is only allowed for the MBs whose corresponding 8x8 block of the base layer is located in an intra coded MB. The prediction residual is transmitted using the H.264/AVC residual coding. In this way, the intra prediction signal is directly obtained by de-blocking and up-sampling the corresponding 8x8 luminance block inside the corresponding lower spatial resolution layer. Therefore, the decoding complexity is significantly reduced, since the inverse MCTF is only required for the MBs that is actually decoded.
- (iii) *Inter-layer residual texture prediction:* The residual information coded in the lower spatial resolution layer is also adaptively used to predict the residual information of the spatial enhancement layer [46]. To achieve this, a flag

indicating the application of residual signal prediction from the lower spatial resolution layer is transmitted for each macro-block. If the flag is true, the base layer residual signals is block-wise up-sampled using a bi-linear filter with constant border extension and used as prediction for the residual signal of the current layer. In this way, only the corresponding difference signal is coded.

2.3.5 Temporal Scalability

The temporal decomposition with hierarchical coding structure permits temporal scalability by removing the information which corresponds to the frames that are not used as reference frame for the subsequent motion estimation and compensation [46].

Figure 2.10: An example of temporal scalability.

An example for the temporal scalability provided by the MCTF based temporal decomposition of a group of 12 frames using 3 decomposition stages is given in Fig. 2.11, where L indicates the temporal low-pass frame, H indicates the

temporal high-pass frames, and M_p indicates the motion information after motion prediction. If all the high-pass frames are discarded and only L^3 is received at the decoder side, the reconstructed video is with $\frac{1}{12}$ temporal resolution of the original input sequence. If the two high-pass frames H^3 are additionally received, the decoder could reconstruct the sequence as L^2 , which is with $\frac{1}{4}$ temporal resolution of the original input sequence. In a same way, the sequence with half of the original temporal resolution could be achieved with further receiving H^2 , and the full temporal resolution sequence could be obtained if H^1 is finally added. Therefore, the low-pass frame L^3 is the temporal base layer, while H^3 , H^2 and H^1 refer to the progressive temporal enhancement layers.

Similar to MCTF, the hierarchical B frame prediction structure can also be employed for supporting several temporal scalability levels. As shown in Fig. 2.7, the key frames (I_0/P_0) represents the coarsest supported temporal resolution, and the temporal resolution could be refined by further including the B frames of the next temporal levels.

2.3.6 Combined Scalability

The wide range of spatial, temporal and SNR scalability can be easily combined into a layered coding scheme. The coding structure depends on the scalability space that is required by the application.

An example of the combined scalable video coding scheme is provided in Fig. 2.11, where L_m^n and H_m^n represent the temporal low-pass and high-pass frames, m is the index of the scalable layer, and n indicates the temporal level (same as in Fig. 2.6). The sequence with QCIF spatial resolution is first encoded at a frame rate of 15Hz as Layer 1. By using the inter-layer prediction method, Layer 2 with CIF spatial resolution is encoded from the up-sampled information of Layer 1. Layer

2.3 Scalable Extension of H.264/AVC

34

Figure 2.11: An example of combined scalability.

3 is achieved based on Layer 2 by the SNR refinement. By additionally encoding the high-pass frames H_4^0 , the temporal frame rate is refined to 30Hz in Layer 4. Similar to Layer 3, Layer 5 is the SNR enhancement layer based on Layer 4.

Chapter 3

Bit Allocation of Rate Control for H.264/AVC

3.1 Introduction

Bit allocation of rate control is an important issue in video streaming applications. If the original video sequence is encoded with fixed quantization parameter (QP), a fairly constant quality reconstructed video could be achieved. However, the bits used for each frame can vary drastically, because the complexity of frames is continually changing in a real video sequence. In the video application, constraints imposed by decoder buffer size and network bandwidth require the encoder to compress the video sequence at a near constant bitrate. For the real-time video communication, a smaller buffer size is used to satisfy the low delay requirement. Here, rate control becomes more challenging because a smaller buffer size can easily cause overflowing and underflowing problems. So, low delay video communication requires more accurate bit allocation strategy.

There are two essential issues that should be considered when designing a bit allocation scheme for rate control. The first issue is about how to assign proper

target bits to each basic unit based on the current network bandwidth and buffer status. The second issue is about how to adjust the encoder parameters (i.e. quantization parameter) when encoding each basic unit to meet the target allocated bits. According to the size of the basic unit, rate control could be implemented at the MB level, the slice level or the frame level [21]. The rate control at smaller basic unit level could achieve more accurate target bits matching and better buffer regulation, at a cost of slight coding efficiency loss.

Based on the classic bit allocation work [53] [54] and combined with appropriate buffer control [55] [56], rate control has been widely studied in digital video coding standards and applications, such as TM5 [14] for MPEG-2 [10], TMN8 [15] and TMN12 [16] for H.263 [11] [17], VM8 [18] [19] for MPEG-4 Visual part 2 [20], JVT-G012 [21] for H.264/AVC [22] and so on. For example, a simple linear rate-distortion model is employed in TM5:

$$R = \frac{K}{Q_s} \quad (3.1)$$

where K is a constant, R is the bits to encode each basic unit, Q_s is the quantization step. It can be seen that a very simplified R-Q model is used in TM5, thus it cannot achieve accurate rate control. An improvement proposed in [57] [58] is based on TM5 by introducing an offset to indicate the overhead bits as:

$$R = \frac{K}{Q_s} + C \quad (3.2)$$

where C is a parameter to specify the non-texture bits. However, the improved TM5 still cannot provide precise rate control due to its simple R-Q model and rough estimation of the offset.

In TMN8, a quadratic R-Q model is used:

$$R = \sigma^2 \times \left(\frac{K}{Q_s^2} + C \right) \quad (3.3)$$

where σ^2 is the variation of the coding coefficients within the basic unit. The TMN8 rate control algorithm is designed for P frames in H.263 video coding. Compared to TM5, the TMN8 could achieve the target bit rate more precisely. However, because there is no boundary for the QP value adjustment at MB level, the TMN8 could suffer large rate control error when scene changes or high motion occurs.

In VM8, the employed R-Q model is:

$$R_{texture} = MAD_{actual} \times \left(\frac{K1}{Q_s^2} + \frac{K2}{Q_s} \right) \quad (3.4)$$

where MAD_{actual} is the actual mean absolute difference of the current basic unit, $K1$ and $K2$ are the second and first order parameters of the R-Q model, and $R_{texture}$ is the target texture bits and derived from

$$R_{texture} = R_{sum} - R_{p_{non-text}} \quad (3.5)$$

with R_{sum} and $R_{p_{non-text}}$ being respectively the target sum bits and the predicted non-texture bits to encode the current basic unit. Note that both σ^2 in TMN8 [15] and MAD in VM8 [18] [19] are used to indicate the coding complexity of the spatial information. In G012 [21] and our scheme, MAD is used as the complexity indicator. The VM8 algorithm is not suitable for H.264/AVC rate control. Due to the way rate distortion optimization (RDO) is used in H.264/AVC [22], the MAD_{actual} is not available before QP calculation. And the $R_{p_{non-text}}$ is very hard to predict in H.264/AVC because of more complicated motion determination.

Besides the above standard rate control algorithms, a rate control scheme with ρ -domain source modelling is proposed in [59–61], where the relation of coding bits and quantization step is indirectly connected with ρ , which is the percentage of zeros among the quantized DCT coefficients. For the transform coding systems, the rate function of the ρ -domain is a linear function. A rate control scheme is proposed based on the ρ -domain model. In [59–61], it was reported that the ρ -domain scheme could achieve accurate and robust rate control and buffer regulation. Besides, the ρ -domain model could estimate the R-D curve before quantization and entropy coding with low complexity. Thus, the ρ -domain based rate control is a potential candidate for real-time visual communication. However, the derivation of the ρ -domain rate control is based on previous standards (MPEG-1, MPEG-2, MPEG-4 Visual part 2, H.261 and H.263), where the RDO is not employed for motion determination. When RDO is used in H.264/AVC, ρ (the percentage of zeros among the quantized DCT coefficients) cannot be estimated before motion estimation. Therefore, it is desired to develop a rate control for the current H.264/AVC video coding standard.

Compared with previous standards of MPEG-1, MPEG-2, MPEG-4 Visual part 2, H.261 and H.263, there are two more challenges for rate control in H.264/AVC:

- (i) Since rate distortion optimization (RDO) is used to search for motion information in H.264/AVC [22], the quantization parameter (QP) affects both the motion estimation and the spatial quantization steps. This means that before the QP value is determined, the actual MAD is not available to indicate the coding complexity of the current frame. On the other hand, MAD is an important variable for determining the QP value. How to adjust QP without the actual MAD is a QP dilemma.
- (ii) With a much more complicated motion estimation strategy adopted by the

H.264/AVC standard [22], a higher percentage of the bits for encoding motion information is required when compared to the previous standards (MPEG-1, MPEG-2, H.261, H.263), especially at low bitrates. And the amount of these non-texture bits fluctuates with hard-to-predict property, so it cannot be simply predicted.

One rate control scheme with temporal MAD prediction method and VM8 [18] [19] based R-Q quadratic model was proposed by Z. G. Li *et al.* in JVT-G012 [21] and was adopted by JVT in the H.264/AVC reference model JM9.8 [62]. In this scheme,

$$R_{texture}[i] = MAD_{pred,temp}[i] \times \left(\frac{K1}{Q_s^2} + \frac{K2}{Q_s} \right) \quad (3.6)$$

and

$$MAD_{pred,temp}[i] = Y1[i] \times MAD_{actual}[i-1] + Y2[i] \quad (3.7)$$

where $MAD_{pred,temp}[i]$ denotes the temporal predicted MAD of the current basic unit in frame i , $MAD_{actual}[i-1]$ denotes the actual MAD of the co-sited basic unit in the temporal previous residual frame, $Y1[i]$ and $Y2[i]$ are the first-order and zeroth-order parameters of this linear prediction model, which would be updated after encoding every basic unit.

However, there are two problems in this existing H.264/AVC rate control scheme:

(a) *Inaccurate MAD prediction:*

If MAD fluctuates due to high motion or scene changes in the test sequence, the linear model performs poorly for such sudden changes. And the updated

model parameters after the abnormal data training can lead to adverse effect in subsequent MAD linear predictions.

(b) *Inaccurate non-texture bit estimation:*

In JVT-G012, the amount of non-texture bits of the current basic unit is simply predicted from the recent encoded basic units, for example by averaging. But actually the amount of non-texture bits fluctuates with unpredictable property and a higher percentage is required than that of the previous standards (MPEG-1, MPEG-2, H.261, H.263), especially at low bitrates. So the error from the simple prediction of non-texture bits will impair the final rate control result.

Although some improvement works of H.264/AVC rate control have been recently published [63–65], the above mentioned two problems are still not solved. If these two problems could be solved properly, improvements and refinements to the rate control scheme for H.264/AVC can be achieved.

The rest of this chapter is organized as follows. We first review the rate distortion optimization (RDO) [23] employed in H.264/AVC standard in Section 3.2. In Section 3.3, we observe the relationship between MAD_{direct} and MAD_{actual} and propose a direct MAD prediction model. A prediction model switching strategy is also introduced to adaptively choose a prediction model with higher accuracy. In Section 3.4, we propose a new sum bits R-Q model to solve the inaccurate texture bit prediction problem. The MB level quantization parameters are optimized in Section 3.5. With the proposed adaptive MAD prediction method and the sum bits R-Q model, our rate control algorithm is described in Section 3.6 and Section 3.7 presents the experimental results and discussions. The concluding remarks are given in Section 3.8.

3.2 Rate Distortion Optimization in H.264/AVC

In the rate control schemes, the statistics of the current predicted residual frame could be used as a measure of coding complexity, such as the variance of residual coefficients used in TMN8 [15] and the mean absolute difference (MAD) used in VM8 [18]. These actual statistical data must be calculated based on determined motion information. In the previous video coding standards (MPEG-1, MPEG-2, H.261, H.263), the motion information is estimated by:

$$mv = \arg \min_{mv \in r} \{D_{mv}\} \quad (3.8)$$

where mv is the determined MB motion vector, r is the motion search range, and D_{mv} is the mean absolute difference between the current original MB and the reference MB. As the QP value is not involved in the motion estimation step, the actual statistical information could be derived independent of the QP value. Therefore, the QP value could be naturally adjusted between the motion compensation step and the spatial quantization step to meet the required coding bits of rate control, as shown in Fig. 3.1.

Figure 3.1: The flowchart of rate control for previous standards (MPEG-1, MPEG-2, H.261, H.263).

3.2 Rate Distortion Optimization in H.264/AVC

42

Figure 3.2: The flowchart of rate control for H.264/AVC.

One of the distinguishing properties of the H.264/AVC standard is its implementation of rate distortion optimization (RDO) to determine motion information [23]. This work is based on the classic rate-distortion theory [66]. With RDO, Lagrangian method is used to provide optimal bit allocation between motion information and residual coefficients. Moreover, in H.264/AVC, the partition of one MB for motion compensation varies from the set {INTRA16x16, INTRA4x4, INTER16x16, INTER16x8, INTER8x16, INTER8x8, INTER8x4, INTER4x8, INTER4x4, SKIP, DIRECT} [42]. Both motion estimation and mode decision should be optimized by using the Lagrangian method [23]. The motion estimation to determine motion vector of every partition is obtained by minimizing:

$$J_{mv} = \{D_{mv} + \lambda_{motion} \times R_{mv}\} \quad (3.9)$$

where R_{mv} denotes the bits to encode the motion vector, and λ_{motion} is the Lagrange multiplier for motion estimation. After the motion vectors for all kinds of motion modes are estimated, the best motion mode is chosen by minimizing:

$$J_{mode} = \{D_{rec} + \lambda_{mode} \times (R_{mv} + R_{coeff})\} \quad (3.10)$$

where D_{rec} is the sum of squared difference (SSD) between the original MB and the reconstructed MB, R_{coeff} is the bits to encode the quantized residual coefficients, and λ_{mode} is the Lagrange multiplier for motion mode decision. The above λ_{motion} and λ_{mode} are decided empirically in [23] as:

$$\lambda_{mode} = 0.85 \times 2^{(Q_p-12)/3} \quad (3.11)$$

$$\lambda_{motion} = \sqrt{\lambda_{mode}} \quad (3.12)$$

Therefore, in H.264/AVC, QP affects both the motion determination and spatial residual quantization. In this way, the statistical data of the residual frame, such as the actual MAD used in VM8, changes with the QP value adjustment as the adjusted QP value also influences the motion information. As shown in Fig. 3.2, QP value adjustment with actual MAD requires the processing of the motion determination multiple of times. This is not desirable in real-time applications as motion determination is the most time consuming part of the whole video coding procedure.

3.3 Improved MAD Prediction

To solve the QP dilemma caused by RDO in H.264/AVC, one simple linear model is proposed in [21] to predict the MAD using previous temporal information, as described in Eq. 3.13. This MAD prediction using previous temporal information is not accurate when MAD changes abruptly due to high motion or scene changes. Moreover, this temporal prediction model is only updated by the available actual MAD data, which means that this linear model only propagates the inherited temporal property of the previous actual MAD. Thus, the updated model is unable to predict current changes, and is less sensitive to input data fluctuations. This is

3.3 Improved MAD Prediction

44

not desirable for a model which is to be used to capture abrupt input changes.

Figure 3.3: The comparison of actual MAD and direct MAD at frame level. Analysis of the P frames from CREW@QCIF-15Hz and FOOTBALL@QCIF-15Hz sequences, IPPP structure with QP=40.

If some other information that is helpful for predicting MAD could be collected before motion determination, then the QP dilemma could be solved. For the frame level rate control (where the basic unit is one frame), after analyzing the original input frames, we find that MAD_{direct} is somewhat related to MAD_{actual} . MAD_{direct} is a measure to evaluate the difference between the current original frame and the previous reconstructed frame, which could be calculated without motion information (as all motion vectors are set to zero). Two example curves of MAD_{direct} and

3.3 Improved MAD Prediction

45

MAD_{actual} at the frame level are shown in Fig. 3.3. It could be observed that the fluctuation of MAD_{direct} always reflects a fluctuation of MAD_{actual} , especially for abrupt changes. Thus, MAD_{direct} is a kind of spatial information of the current frame, and could help in the prediction of MAD_{actual} through another spatial linear prediction model:

$$MA \cap_{red,spat}[i] = Z1[i] \times MAD_{direct}[i] + Z2[i] \quad (3.13)$$

where $MAD_{pred,spat}[i]$ is the spatial predicted MAD of the current basic unit in frame i , $MAD_{direct}[i]$ represents the direct MAD of the current basic unit in frame i , $Z1[i]$ and $Z2[i]$ are the first-order and zeroth-order parameters of this linear prediction model, which would be updated after encoding every basic unit.

Although the above spatial MAD prediction could efficiently predict the abrupt changes of MAD, for some sequences with fine texture details or irregular regional motion, this spatial prediction model sometimes does not work as well at the frame level. Therefore, we introduce the following two similarity measures to indicate the efficiencies of temporal MAD prediction model and spatial MAD prediction model:

$$\Gamma_{temp}[i] = \sum_{n=i-S}^i |MAD_{pred,temp}[n] - MAD_{actual}[n]| \quad (3.14)$$

$$\Gamma_{spat}[i] = \sum_{n=i-S}^i |MAD_{pred,spat}[n] - MAD_{actual}[n]| \quad (3.15)$$

where S is the number of MAD samples used to measure Γ . The method to predict MAD could then be adaptively switched between the spatial model and the temporal model:

if $\Gamma_{spat}[i] > \Gamma_{temp}[i]$ **then**

$$MAD_{pred,adapt}[i+1] = MAD_{pred,temp}[i+1]$$

3.3 Improved MAD Prediction

46

else

$$MAD_{pred,adapt}[i+1] = MAD_{pred,spat}[i+1]$$

end if

The comparison of $MAD_{pred,temp}$ and $MAD_{pred,adapt}$ is given in Fig. 3.4. Meanwhile, the prediction errors of different models and the percentage of temporal model in use are measured at the frame level. The details are shown in Table 3.1.

Figure 3.4: The comparison of actual MAD, linear predicted MAD and proposed adaptively predicted MAD at frame level. Analysis of the P frames from CREW@QCIF-15Hz and FOOTBALL@QCIF-15Hz sequences, respectively. IPPP structure with QP=40.

From Fig. 3.4 and Table 3.1, it is clear that our adaptive model could adaptively switched between the temporal model and spatial model to achieve a more accurately predicted MAD. Compared with the temporal MAD prediction model

3.3 Improved MAD Prediction

47

Table 3.1: The Percentage of Temporal Model in Use and The Prediction Error of Different Models at Frame Level. IPPP structure with QP=40.

Sequence QCIF 15Hz	temp. predictor in use	Prediction Error		
		temp.	spat.	adapt.
TABLE	19.6%	0.49 (7.3%)	0.24 (3.6%)	0.22 (3.2%)
FOOTBALL	16.2%	0.11 (0.8%)	0.15 (1.2%)	0.10 (0.8%)
BUS	51.1%	0.77 (11.8%)	0.59 (9.1%)	0.59 (9.1%)
CREW	23.3%	0.21 (2.0%)	0.34 (3.2%)	0.19 (1.8%)

proposed in JVT-G012 [21], our adaptive model is significantly better, especially for the sequences with high motion or scene changes.

Similar idea can also be applied to the MB level MAD prediction. Here, the MAD_{direct} could be calculated based on neighboring predicted motion information, which is easily available without motion estimation. Because the motion vectors for neighboring partitions are often highly correlated, the motion vector predicted from nearby vectors is often close to the final determined motion information. So in the MB level rate control, MAD_{direct} derived based on predicted motion vector is more correlated to MAD_{actual} than MAD_{direct} derived without any motion information, and in this way MAD_{direct} could reflect more regional properties such as regional high motion and fine texture. The similarity measurement could also be employed to adaptively choose the more efficient MB level MAD prediction scheme, as described before.

Fig. 3.5 compares our adaptive MAD prediction model and the temporal MAD prediction model at the MB level. It is clear that our proposed scheme is better at predicting regional abrupt changes due to regional high motion and fine texture.

In summary, our proposed MAD prediction model can accurately estimate the statistical property before motion estimation, especially when high motion or scene change happens. Our MB level adaptive MAD prediction further improves the prediction efficiency when regional high motion or fine texture occurs.

3.3 Improved MAD Prediction

48

Figure 3.5: The comparison of actual MAD, linear predicted MAD and proposed adaptively predicted MAD at MB level. Randomly analyzed the 67th frame from HARBOUR@QCIF-15Hz, 90th frame from FOREMAN@QCIF-15Hz, 35th frame from FOOTBALL@QCIF-15Hz and 5th frame from CREW@QCIF-15Hz sequences, with QP=40.

3.4 Sum Bit R-Q model

In the rate control schemes for the previous standards (MPEG-1, MPEG-2, H.261, H.263) [14–16, 18, 19], the amount of non-texture bits of the current basic unit is simply predicted from the recent encoded basic units, for example by averaging. So, the amount of texture bits is derived as Eq. (3.5). In previous standards, the amount of bits used for non-texture encoding is much less than the texture bits, and the amount of non-texture bits varies smoothly. So the prediction error of Eq. (3.5) is still acceptable. However, the H.264/AVC standard newly introduces many complicated motion estimation modes [22], which leads to some increase of the non-texture bits. At the same time, due to the development of more efficient motion estimation, the energy in the residual frame is reduced. Therefore, non-texture bits occupy a higher percentage of the sum bits, especially at low bitrates. And there also exists hard-to-predict fluctuations in the non-texture bits, as shown in Fig. 3.6. So Eq. (3.5) might not be suitable for estimating the amount of texture bits for H.264/AVC. The prediction of non-texture bits with fluctuation is a very complicated task, and the error of such prediction will finally yield a poor rate control result.

To solve this problem, we found that the process of predicting the amount of non-texture bits might be abridged if one model could be formulated to directly relate the amount of sum bits to the QP value. From empirical results, the MAD could be used as a complexity indicator to encode one frame into sum bits. The linear relationship between sum bits and MAD could also be observed in Fig. 3.7. To construct an R-Q model, one approach is to analyze the structure of the video processing and the statistical properties of the video data. This is called the analytical approach. Another approach is to derive the R-Q relationship based on several sampled values, which is called the empirical approach. Our proposed

3.4 Sum Bit R-Q model

50

Figure 3.6: The variation of the amount of texture bits and non-texture bits. Analysis of HARBOUR@CIF-15Hz, FOREMAN@QCIF-15Hz, FOOTBALL@QCIF-15Hz and CREW@QCIF-15Hz sequences, with QP=40.

approach is based on the result in [14] and [21], which belongs to the analytical approach, but we make some improvements according to the empirical result.

Figure 3.7: The relationship between sum bit and MAD with fixed QP=40. Analysis of FOOTBALL@QCIF-15Hz and CREW@QCIF-15Hz sequences.

Because the QP value affects both motion and texture information in H.264/AVC, it is not sufficient to just consider the relationship between QP value and texture bits. One example of the relationship among MAD, QP and texture/non-texture/sum bits is shown in Fig. 3.8. It can be observed that both $\frac{TextBit}{MAD}$ and $\frac{Non-TextBit}{MAD}$ are monotonic increasing with $\frac{1}{Q_p}$, so their summation should also be a monotonic increasing function. The relation between $\frac{SumBit}{MAD}$ and $\frac{1}{Q_p}$ is also shown in Fig. 3.8. The problem here is to find a mathematical model to approximate this monotonic increasing relation between $\frac{SumBit}{MAD}$ and $\frac{1}{Q_p}$. One can employ either the quadratic or the linear model:

$$R_{sum}[i] = MAD_{pred,adapt}[i] \times \left(\frac{X1[i]}{Q_p[i]} + \frac{X2[i]}{Q_p[i]^2} \right) \quad (3.16)$$

$$R_{sum}[i] = MAD_{pred,adapt}[i] \times \frac{X1[i]}{Q_p[i]} + X2[i] \quad (3.17)$$

3.4 Sum Bit R-Q model

52

where $R_{sum}[i]$ is the target sum bit of the basic unit to be encoded, which could be calculated according to the buffer status and the remaining bits, $X1[i]$ and $X2[i]$ are two parameters of the R-Q model, that could be updated with linear regression method after encoding each frame, and $Q_p[i]$ is the QP value.

Table 3.2: Comparison of R-Q Model Accuracy. IPPP structure with initial QP=40.

Sequence QCIF 15Hz	Target Bits	Mismatch		
		G012	Ours, quad	Ours, linear
CREW	2000	638	471	475
CITY	800	186	102	124
MOBILE	2266	524	366	359

Mathematically, the quadratic model could achieve more accurate approximation. One simulation is carried out to test the R-Q model accuracy, and the result is given in Table 3.2. In this simulation, all the three R-Q models are used at the frame level, the target bits for each frame is set to a fixed value. All the R-Q models are updated with linear regression method after encoding each frame. The sum bits and adaptive MAD prediction methods are employed in our quadratic and linear models as in Eqs. (3.16) and (3.17). From Table 3.2, both our quadratic and linear R-Q models with sum bits and adaptive MAD prediction outperform the G012 model (Eqs. (3.6) and (3.13)). The quadratic and linear models have comparable approximation accuracies, but the QP calculation and the model training process of quadratic model are more complicated. If MB level rate control is activated, the QP calculation and model training of quadratic model for each MB will significantly increase the encoding complexity. To trade-off the complexity and accuracy, linear model is chosen in our scheme. Compared to the G012 model, the approximation error is reduced by up to 33% with our linear model. The accuracy of the linear approximation is also specified in Fig. 3.8 by the R-squared value

$(\mathbb{R}^2)$:

$$\mathbb{R}^2 = 1 - \frac{SSE}{SST} \quad (3.18)$$

where

$$SSE = \sum (\mathbb{Y}_i - \hat{\mathbb{Y}}_i)^2 \quad (3.19)$$

and

$$SST = \left(\sum \mathbb{Y}_i^2 \right) - \frac{(\sum \mathbb{Y}_i)^2}{n} \quad (3.20)$$

$\mathbb{R}^2$ is an indicator from 0 to 1 that reveals how closely the approximated linear function ($\hat{\mathbb{Y}}_i$) is to the actual data ($\mathbb{Y}_i$). The approximated linear function is most reliable when its R-squared value is at or near 1. SSE stands for the sum-of-squares error, and SST is the sum-of-squares total. Note that, in H.264/AVC, the relationship between QS (quantization step) and QP is logarithmic as shown in Eq. (3.24). QP values of 20 to 40 correspond to QS values of 6.5 to 64. So the bitrate at QP=20 is almost 10 times that of the bitrate at QP=40. This bitrate range is sufficient to cover the video quality from high to low. The QP values outside this range are seldom used for real time video communication. Therefore, in Fig. 3.8, the relationship between $\frac{SumBit}{MAD}$ and $\frac{1}{Q_p}$ is only shown for QP value that ranges from 20 to 40, and the QP values of this range are widely used in this research area [65, 67, 68]. Moreover, QP value is usually adjusted within a small range for the successive frames to maintain a smooth visual quality. Clearly, the proposed linear model can also work properly within such a small range, even with very high or very low QP values.

3.4 Sum Bit R-Q model

54

Therefore, with the linear R-Q model of Eq. (3.17), $Q_p[i]$ could be calculated based on $R_{sum}[i]$ and $MAD_{pred,adapt}[i]$, which are both available before motion estimation.

Figure 3.8: The linear relationship among TextureBit, Non-TextureBit, SumBit and QP. Randomly chosen 1st and 2nd P frames from FOOTBALL@QCIF-15Hz sequence, QP=20:2:40 (QP values are set from 20 to 40, with increase step of 2).

A similar linear R-Q model has been used in some rate control work [65] based on TM5 [14] for MPEG-2 rate control:

$$R_{sum}[i] = MAD_{actual}[i] \times \frac{K}{Q_s[i]} + C \quad (3.21)$$

where K and C are two constants and could be calculated based on the texture

bits and non-texture bits after encoding each frame. The zeroth-order parameter C is the average bits to encode the motion information and other header bits, and it is a kind of simple predicted value of non-texture bits, which is not suitable for the H.264/AVC standard.

Compared with Eq. (3.21), there are three differences in our proposed model:

- (i) $X2[i]$ in our proposed sum bits R-Q model denotes the prediction residual. Here, $X1[i]$ and $X2[i]$ are not related to the texture bits and non-texture bits. They could be updated by a linear regression method after encoding each frame. In this way, the inaccurate estimation for the amount of non-texture bits is abridged. The final rate control result is improved because the negative effect of inaccurate texture bits estimation is eliminated;
- (ii) “Q” in our proposed R-Q model is the quantization parameter instead of quantization step in Eq. (3.21). So the calculation procedure between quantization parameter and quantization step is also abridged. This reduces the computing complexity.
- (iii) MAD, which is used to specify the coding complexity, is a predicted value in our sum bits R-Q model, instead of actual value in Eq. (3.21).

3.5 Optimized QP Adjustment at MB level

It is known that, for a zero-mean i.i.d. source, the distortion (D) versus uniform quantization step relation could be approximated as follows [69] [54]:

$$D(Q_s) = \frac{Q_s^2}{\varepsilon} \quad (3.22)$$

3.5 Optimized QP Adjustment at MB level

56

where Q_s is the quantization step and ε is a source dependent parameter.

In the previous standards, the relationship between quantization step and quantization parameter is always linear:

$$Q_s = 2 \times Q_p \quad (3.23)$$

H.264/AVC has newly proposed integer transform and division free quantization [70], and the relationship between QS and QP becomes

$$Q_s = 2^{Q_p/6} \times \nu(Q_p \text{ mode } 6) \quad (3.24)$$

where

$$\begin{aligned} \nu(0) &= 0.675 \\ \nu(1) &= 0.6875 \\ \nu(2) &= 0.8125 \\ \nu(3) &= 0.875 \\ \nu(4) &= 1.0 \\ \nu(5) &= 1.125 \end{aligned}$$

The relation between QS and QP could be accurately approximated as an exponential function, as shown in Fig. 3.9:

$$Q_s = C_1 \times C_2^{Q_p} \quad (3.25)$$

where C_1 and C_2 are two constants. According to Eqs. (3.22) and (3.25), if we use

Figure 3.9: The relation of QP and QS in H.264/AVC.

mean square error (MSE) to indicate the $D(Q_s)$, the PSNR could be derived as

$$\begin{aligned}
 PSNR &= 10 \log_{10} \frac{255^2}{MSE} \\
 &= 10 \log_{10} \frac{255^2}{\frac{Q_s^2}{\epsilon}} \\
 &= 10 \log_{10} \frac{\epsilon * 255^2}{(C_1 \times C_2^{Q_p})^2} \\
 &= a * QP + b
 \end{aligned} \tag{3.26}$$

where a and b are two constants. This linear relation of QP and PSNR could be observed in Fig. 3.10.

At MB level rate control, some MBs within one region may share the same QP value to reduce the extra bits to encode QP differences, one region may contain several continuous MBs which are connected to each other horizontally or vertically. To obtain reconstructed video quality over the whole frame, the PSNR value of the whole frame should be maximized with given target bits, as

$$\max \sum_{i=1}^N [(a \times Q_p[i] + b) * w_i] \tag{3.27}$$

where N is the number of regions within each frame, and w_i is the number of MBs

3.5 Optimized QP Adjustment at MB level

58

Figure 3.10: The linear relationship between PSNR and QP. Randomly chosen 1st, 2nd, 3rd and 4th P frames from FOOTBALL@CIF-15Hz sequence, QP=20:2:40.

3.6 Proposed Rate Control Algorithm

59

within the i -th region. According to Eq. (3.17), the sum bits constraint is

$$R_{sum} = \sum_{i=1}^N [(MAD_{pred,adapt}[i] * \frac{X1[i]}{Q_p[i]} + X2[i]) * w_i] \quad (3.28)$$

The classic Lagrange theory [71] can be used to obtain the optimized QP value for each region:

$$\begin{aligned} Q_p[i] &= \frac{X1[i]}{R_{sum} - X2[i] \sum_{k=1}^N w_k} \sqrt{MAD_{pred,adapt}[i]} \\ &\quad * \sum_{k=1}^N (\sqrt{MAD_{pred,adapt}[k]}) \end{aligned} \quad (3.29)$$

For simplicity, the number of MBs within one region is set to 1 in our scheme, that means $w_i = 1, 1 \leq i \leq N$. Then Eq. (3.29) becomes

$$\begin{aligned} Q_p[i] &= \frac{X1[i]}{R_{sum} - NX2[i]} \sqrt{MAD_{pred,adapt}[i]} \\ &\quad * \sum_{k=1}^N (\sqrt{MAD_{pred,adapt}[k]}) \end{aligned} \quad (3.30)$$

where N is now the number of MBs in one frame. Eq. (3.30) would be used to adjust QP at MB level in our rate control scheme, which will be described in details in the next Section.

3.6 Proposed Rate Control Algorithm

With the adaptive MAD prediction method and the sum bits R-Q model, the detailed block diagram of our rate control scheme at the frame level is shown in Fig. 3.11. There are essentially five stages for frame level rate control: frame level MAD prediction, frame level target bit allocation, frame level QP adjustment, MB

3.6 Proposed Rate Control Algorithm

60

Figure 3.11: The block diagram of our rate control scheme at frame level.

3.6 Proposed Rate Control Algorithm

61

level rate control, parameter updates of frame level sum bits R-Q model and MAD prediction model.

1. *Frame level MAD prediction:*

The encoder adaptively chooses the MAD prediction model with higher prediction efficiency from either temporal or spatial MAD prediction model, as described in Section 3.3.

2. *Frame level target bit allocation:*

The target bit to encode each frame is calculated from

$$R_{sum}[i] = (1 - \alpha) \times R_1[i] + \alpha \times R_2[i] \quad (3.31)$$

where $R_{sum}[i]$ is the target sum bits to encode the i -th frame, including texture, motion and other header information; α is a constant that is set to 0.5 in JVT-G012 [21];

$$R_1[i] = \frac{W}{f} + \beta \times (B_t - B_c) \quad (3.32)$$

i.e. $R_1[i]$ is estimated from the previous actual buffer occupancy B_c , target buffer occupancy B_t , frame rate f and the available bandwidth W , with β set to 0.75 in JVT-G012 [21] with concern of tight buffer regulation;

$$R_2[i] = \gamma \times \frac{R_r[i]}{N_r[i]} + (1 - \gamma) \times \sigma \times \overline{R_{act}}[i - 1] \quad (3.33)$$

i.e. $\overline{R_{act}}[i - 1]$ is the average value of actual sum bits used to encode the previous frame, $R_2[i]$ is adaptively assigned according to the remaining bits and frame complexity. In Eq. (3.32), $R_r[i]$ and $N_r[i]$ are the amount of remaining bits and the

3.6 Proposed Rate Control Algorithm

62

number of remaining frames in the current GOP before encoding the i -th frame:

$$R_r[i] = \begin{cases} \frac{W}{f} \times GOP_{size} & : i = 1 \\ R_r[i-1] - R_{act}[i-1] & : i > 1 \end{cases} \quad (3.34)$$

$$N_r[i] = \begin{cases} GOP_{size} & : i = 1 \\ N_r[i-1] - 1 & : i > 1 \end{cases} \quad (3.35)$$

GOP_{size} is the size of GOP, γ is a weighting factor that is set to 0.875 in our experiments, and σ is a variable to specify the ratio of coding complexity between the current frame and previously encoded frame. σ is introduced with the aim to adaptively allocate more bits to encode the frame with more information, and it is given by

$$\sigma = \frac{MAD_{pred,adapt}[i]}{\overline{MAD_{actual}[i-1]}} \quad (3.36)$$

where $\overline{MAD_{actual}}$ is the average value of previous MAD_{actual} . σ is further restrained as

$$\sigma = \begin{cases} 1 & : 0.5 < \sigma < 1.5 \\ \sigma & : \text{else} \end{cases} \quad (3.37)$$

Eq. (3.37) aims to activate the impact of σ only when a sudden MAD fluctuation happens, e.g. caused by a real scene change or motion change. An example to show the effect of σ is given in Section 3.7.

Finally, with the hypothetical reference decoder (HRD) which has been defined

3.6 Proposed Rate Control Algorithm

63

in H.264/AVC [72], the target sum frame bits are bounded by

$$R_{sum}[i] = \max\{L_{HRD}[i], \min\{R_{sum}[i], U_{HRD}[i]\}\} \quad (3.38)$$

where $L_{HRD}[i]$ and $U_{HRD}[i]$ are respectively the lower and upper bounds of HRD for the current frame, they could be calculated as

$$L_{HRD}[i] = \begin{cases} R_{preGOP} + \frac{W}{f} & : i = 1 \\ L_{HRD}[i-1] + (\frac{W}{f} - R_{act}[i-1]) & : i > 1 \end{cases} \quad (3.39)$$

$$H_{HRD}[i] = \begin{cases} (R_{preGOP} + B_{full}) \times \rho & : i = 1 \\ H_{HRD}[i-1] + (\frac{W}{f} - R_{act}[i-1]) \times \rho & : i > 1 \end{cases} \quad (3.40)$$

where R_{preGOP} is the difference between the target bits and actual bits of the previous GOP, B_{full} is the full buffer size, which is determined by the current bandwidth W and minimal buffer delay time t_{delay} :

$$B_{full} = W \times t_{delay} \quad (3.41)$$

where ρ is a constant with the value of 0.9.

3. Frame level QP adjustment:

When the target sum bit to encode the i -th frame is determined, the QP value could be calculated by Eq. (3.17). Since the inaccurate estimation of the amount of texture bits is abridged by our refined R-Q model proposed in Section 3.4, the adverse effect caused by the rough setting of non-texture bits as a predicted value is eliminated.

To maintain the smoothness of visual quality within one sequence, the QP value

3.6 Proposed Rate Control Algorithm

64

is further adjusted by

$$Q_p[i] = \min\{\overline{Q_p}[i-1] + 2, \max\{\overline{Q_p}[i-1] - 2, Q_p[i]\}\} \quad (3.42)$$

where $Q_p[i]$ is the QP value for the i -th frame, and $\overline{Q_p}[i-1]$ is the average QP value for all MBs in the previous frame.

The adjusted QP value derived from Eq. (3.42) is applied as the initial QP to the MB level rate control of current frame.

4. MB level rate control:

The MB level rate control could be viewed as an optional function in our scheme. If the size of the basic unit is set as one frame, the MB level rate control is skipped by directly using the frame level predicted QP to encode all MBs within the frame. For low delay application with small buffer size, the MB level rate control is always activated as it could help the encoder to achieve more accurate target bit matching and buffer control at the cost of minor coding efficiency loss.

MB level rate control could be further enhanced by another five steps including: MB level remaining bits calculation, MB level MAD prediction, MB level QP adjustment, MB actual encoding, parameters update of MB level sum bits R-Q model and MAD prediction model. The detailed block diagram of our rate control scheme at the MB level is illustrated in Fig. 3.12.

4.1 MB level remaining bits calculation:

The remaining bits of MB level rate control is initialized as R_{sum} , which is the amount of target bits for current frame. For the following MBs, T is calculated at the subsequent of MB level rate control for each MB:

$$T[i] = \begin{cases} R_{sum} & : i = 0 \\ T[i-1] - R_{mb}[i-1] & : \text{else} \end{cases} \quad (3.43)$$

3.6 Proposed Rate Control Algorithm

65

Figure 3.12: The block diagram of our rate control scheme at MB level.

3.6 Proposed Rate Control Algorithm

66

where $R_{mb}[i - 1]$ is the actual sum bits used to encode previous MB.

4.2 MB level MAD prediction:

At the MB level, the MAD_{direct} is calculated based on the motion information predicted from spatial neighboring MBs, as described in Section 3.3. Note that the spatial correlation which could be used to predict motion information is only available for the current MB. For all other remaining MBs, the predicted motion vectors are zero, so their MAD_{direct} should be calculated in the same way as the frame level. Our $MAD_{pred,adapt}$ always yields a more accurate MAD prediction because of the switching policy as described in Section 3.3.

Sometimes when the encoding sequence is characterized with high regional motion, such as *FOOTBALL*, a rough motion estimation could be activated as an optional function to obtain more accurate MAD prediction. At the start of MB level rate control, INTRA16, INTER16 and the most possible mode could be utilized to achieve mode decision with the predicted frame QP for all the MBs within the current frame, here the most possible mode is set to the determined motion mode of the co-site MB within the previous frame due to similarity of the temporal motion. The information of these three modes are saved at the encoder side, and the derived MAD is assigned to $MAD_{pred,adapt}$. With this rough motion estimation, $MAD_{pred,adapt}$ becomes more related to the MAD_{actual} . In the following motion determination step, these three modes of rough motion estimation are not carried out anymore, their information could be obtained directly from previously saved data, so this rough motion estimation would not incur more computing complexity. The only expense for this method is that the three modes are first processed with fixed QP for all MBs, but other modes are with adjusted QP. However, this effect could be neglected when the QP adjustment varies within a small boundary.

4.3 MB level QP adjustment:

3.6 Proposed Rate Control Algorithm

67

If the current MB is the first one in the current frame, its QP value is directly set to the frame level predicted QP.

If the remaining bits before encoding the current MB $T[i] < 0$, the QP value is increased by 1 to achieve frame level actual bits that are closer to target bits, i.e.

$$Q_p[i] = Q_p[i - 1] + 1 \quad (3.44)$$

Otherwise, the MB level QP could be adjusted according to Eq. (3.45):

$$Q_p[i] = \frac{X1[i]}{T[i] - (N - i)X2[i]} \sqrt{MAD_{pred,adapt}[i]} * \sum_{k=i}^N (\sqrt{MAD_{pred,adapt}[k]}) \quad (3.45)$$

The derived QP value should also be restricted by the QP value of the previously encoded MB to reduce blocking artifacts:

$$Q_p[i] = \min\{Q_p[i - 1] + 1, \max\{Q_p[i - 1] - 1, Q_p[i]\}\} \quad (3.46)$$

Moreover, the boundary to maintain visual smoothness along the temporal direction should also be considered, i.e.

$$Q_p[i] = \min\{\overline{Q_p} + 3, \max\{\overline{Q_p} - 3, Q_p[i]\}\} \quad (3.47)$$

where $\overline{Q_p}$ specifies the average QP value for all MBs in the previous frame.

Finally, the QP value should be restricted between 0 and 51, which is provided by H.264/AVC [42]:

3.6 Proposed Rate Control Algorithm

68

$$Q_p[i] = \min\{51, \max\{0, Q_p[i]\}\} \quad (3.48)$$

4.4 MB actual encoding:

The final adjusted QP value derived from Eq. (3.48) could be applied to motion estimation with RDO to the current MB. The QP value is also used to derive QS which is applied to quantize the spatial data before entropy coding. The relationship between QP and QS of H.264/AVC is specified in Eq. (3.24).

4.5 Parameters updates of MB level sum bits R-Q model and MAD prediction model:

After encoding each MB, the encoder should update the parameters of the MB level sum bits R-Q model and the linear MAD prediction model.

First, one sliding-window mechanism [18] is employed to adaptively choose the previous H MBs to update the model as sample data. The sliding-window size H depends on the fluctuation extent of coding complexity (MAD_{actual}). For example, if the coding complexity changes abruptly due to scene changes or high motion, H should be adaptively set to a smaller value so that less recent data is used. This sliding-window mechanism could be mathematically expressed as:

If $MAD_{actual}[i-1] \geq MAD_{actual}[i]$,

$$H = H_{max} \times \left(\frac{MAD_{actual}[i]}{MAD_{actual}[i-1]} \right) \quad (3.49)$$

else,

$$H = H_{max} \times \left(\frac{MAD_{actual}[i-1]}{MAD_{actual}[i]} \right) \quad (3.50)$$

3.6 Proposed Rate Control Algorithm

69

where H_{max} is the maximal sliding-window size predefined as 20 in our experiments.

$MAD_{actual}[i]$ is the actual MAD of the i -th MB of the current frame.

After encoding each MB, the first-order and zeroth-order parameters ($X1[i]$, $X2[i]$) of the R-Q model could be updated with the recent H sample data using linear regression method. To reduce the negative impact of some erroneous data points during the R-Q model update process, some poor data should be further removed as outliers. The amount of target bits could be recalculated by the first updated R-Q model with given actual QP value, and some statistical properties (such as standard deviation) between the amount of actual bits and recalculated target bits could be used as an indicator to reject poor data. Therefore, only representative data from the recent H candidates are selected to finally update the R-Q model. In this way, the R-Q model could be further calibrated by introducing the rejection criterion of outliers.

Similar parameter update strategy with outlier removal is also employed to update the first-order and zeroth-order parameters ($Y1[i]$, $Y2[i]$) of the traditional linear MAD prediction model. Note that there is no parameter update for our proposed direct MAD prediction model.

5. Parameters updates of frame level sum bits R-Q model and MAD prediction model:

The parameters of frame level sum bits R-Q model and linear MAD prediction model could also be updated after the encoding of each frame. This procedure is similar to the MB level parameter update, except that the training sample data is from the frame level encoding.

All the parameters mentioned in our scheme are listed in Table 3.3 and 3.4 for the frame level and the MB level rate control respectively. Some parameters are predefined heuristically as constants. For example, α is a weighing factor to

3.6 Proposed Rate Control Algorithm

70

Table 3.3: The parameters used in the frame level rate control.

parameters	description	default value
i	the index of frame	-
R_{sum}	the target sum bits to encode one frame	-
R_1	the bits estimated from buffer occupancy	-
R_2	the bits estimated from remaining bits and frame complexity	-
α	a constant	0.75
β	a constant	0.5
W	the available bandwidth	-
f	the frame rate	-
B_t	the target buffer occupancy	-
B_c	the actual buffer occupancy	-
γ	a constant	0.875
R_r	the amount of the remaining bits	-
N_r	the number of the remaining frames	-
σ	the parameter to indicate relative frame complexity	-
$\overline{R_{act}}$	the average bits to encode previous frames	-
GOP_{size}	the size of GOP	-
$\overline{MAD_{actual}}$	the average actual MAD of previous frames	-
L_{HRD}	the lower HRD bound	-
H_{HRD}	the higher HRD bound	-
R_{preGOP}	the remaining bits of previous GOP	-
B_{full}	the full buffer size	-
t_{delay}	the buffer delay time	-
ρ	a constant	0.9
Q_p	the initial QP of the MB level rate control	-
$\overline{Q_p}$	the average MB QP value within one frame	-

Table 3.4: The parameters used in the MB level rate control.

parameters	description	default value
i	the index of MB	-
T	the remaining bits for current frame	-
R_{mb}	the bits used for one MB	-
Q_p	the calculated QP for the MB	-
H	the window size of update	-
H_{max}	the maximal window size	20

balance R_1 and R_2 , and it is set to 0.5. β is constant to weight the buffer mismatch, and it is set to 0.75 for tight buffer regulation. Other reasonably different setting of these parameters will not affect the final results significantly.

Besides the advantages of adaptive MAD prediction model and sum bits R-Q model which we already described in Sections 3.3 and 3.4, there are two other improvements in our rate control strategy as compared to JVT-G012 [21]:

- (a) σ defined by Eq. (3.36) and (3.37) are newly introduced with the aim to adaptively allocate more bits to encode the frame with higher MAD. It improves the subjective video quality for the frame with more information.

- (b) The frame level predicted QP value is used as the initial QP at the MB level rate control, instead of using the average QP of the previous frame. This ensures that the MB level rate control starts from a closer range to meet the target frame bits.

3.7 Experimental Results

The performance of our proposed rate control scheme for H.264/AVC is evaluated in this section. The simulation is implemented with the JVT reference software JM9.8 [62], in which JVT-G012 [21] is adopted to achieve rate control. To compare our scheme with JVT-G012 for real-time application, the MB level rate control is activated for both schemes, all the test sequences are intra-coded for the first frame (I frame) and followed with subsequent inter-coded frames (P frames).

Table 3.5: The Comparison of the Target Frame Bits Mismatch.

Sequence	FR RC			MB RC		
	G012	Ours	Red.	G012	Ours	Red.
CREW	30.2%	22.0%	8.2%	19.4%	10.1%	9.3%
FOOTBALL	16.4%	13.4%	3.0%	13.5%	9.0%	4.5%
HARBOUR	34.1%	22.4%	11.7%	26.2%	6.6%	19.6%
FOREMAN	55.4%	35.8%	19.6%	43.7%	15.2%	28.5%

We first compare the percentage of target frame bits mismatch in Table 3.5 with frame skipping being turned off, as the comparison of target bits mismatch for the two schemes with different number of skipped frames is unfair. The comparison is implemented in both the frame level rate control (“FM RC” in Table 3.5) and the MB level rate control (“MB RC” in Table 3.5) for four sequences. The target bits mismatch is reduced by up to 28% with our scheme. Fig. 3.13 shows this comparison frame by frame at the MB level rate control for the sequences *FOREMAN* and *HARBOUR*. Fig. 3.14 presents the corresponding PSNR plots to Fig. 3.13.

3.7 Experimental Results

72

Figure 3.13: The comparison of target frame bits mismatch between our scheme and JVT-G012 at MB level rate control. Analysis of FOREMAN@QCIF-15Hz and HARBOUR@QCIF-15Hz sequences, with sequence initial QP=40.

3.7 Experimental Results

73

Figure 3.14: The comparison of PSNR between our scheme and JVT-G012 at MB level rate control. Analysis of FOREMAN@QCIF-15Hz and HARBOUR@QCIF-15Hz sequences, with sequence initial QP=40.

3.7 Experimental Results

74

Figure 3.15: The comparison of buffer status between different coding scheme at MB level rate control. Analysis of CITY@QCIF-15Hz, FOREMAN@QCIF-15Hz, BUS@QCIF-15Hz and MOBILE@QCIF-15Hz sequences, with sequence initial QP=40.

3.7 Experimental Results

75

Table 3.6: The Comparison of Final Rate Control Result.

Sequence	RC Sch.	Bitrate (kb/s)	Frame Sk.	Bit Was.	Tgt. bit Mism.	PSNR (dB)
FOREMAM QCIF 15Hz	G012	19.06	15	0	468	27.55
	Ours	19.02	4	0	237	27.81
	Ours #	19.04	4	0	210	27.86
FOOTBALL QCIF 15Hz	G012	63.87	27	4253	641	26.01
	Ours	63.98	21	4321	440	26.27
	Ours #	63.98	8	4110	340	27.14
HABOUR QCIF 15Hz	G012	20.07	8	0	300	24.54
	Ours	20.02	8	0	142	24.84
	Ours #	20.02	8	0	112	24.95
SOCCER QCIF 15Hz	G012	32.09	3	0	458	28.74
	Ours	32.07	1	0	193	28.85
	Ours #	32.07	1	0	172	28.95
TABLE QCIF 15Hz	G012	32.03	5	40	661	30.69
	Ours	32.03	4	0	344	30.92
	Ours #	32.03	4	0	331	31.03
CREW CIF 15Hz	G012	96.04	2	420	753	30.10
	Ours	96.04	1	488	376	30.05
	Ours #	96.02	1	431	355	30.08
CITY CIF 15Hz	G012	40.07	10	0	852	26.73
	Ours	40.04	10	0	473	26.81
	Ours #	40.04	10	0	451	26.85
ICE QCIF 15Hz	G012	30.14	5	0	552	29.49
	Ours	30.03	3	0	139	29.83
	Ours #	30.03	3	0	137	29.89
MOBILE QCIF 7.5Hz	G012	21.95	7	372	839	22.90
	Ours	21.95	7	337	824	23.05
	Ours #	21.97	7	334	528	23.14
BUS QCIF 7.5Hz	G012	25.82	8	668	524	24.57
	Ours	25.82	8	597	367	24.66
	Ours #	25.82	8	532	288	24.76

3.7 Experimental Results

76

Figure 3.16: The comparison of reconstructed video quality, the 77th, 78th, 79th and 80th frames of FOREMAN. (a) reconstructed with our scheme, (b) reconstructed with JVT-G012.

Figure 3.17: The comparison of reconstructed video quality, the 62th, 63th, 64th and 65th frames of FOOTBALL. (a) reconstructed with our scheme, (b) reconstructed with JVT-G012.

3.7 Experimental Results

77

Figure 3.18: The comparison of reconstructed video quality. (a) reconstructed without σ effect, (b) reconstructed with σ effect, (c) details of (a), (d) details of (b).

For real-time video communication, the buffer size is set as $0.1 * \text{bitrate}$. In other words, the maximal buffer delay is limited to 100ms to satisfy the low delay requirement. When the buffer is full, the encoder skips frames until there is available space in the buffer. When the buffer is empty and the input bit to the buffer is smaller than the output bandwidth, buffer underflowing will cause low channel utilization. The major objectives of our rate control scheme are to regulate the amount of sum bit to encode each P frame, efficiently utilize the buffer resource and prevent the buffer from overflowing and underflowing. Due to our proposed accurate bit allocation scheme, the buffer is well controlled. The buffer fullness status is compared in Fig. 3.15. Our scheme in Fig. 3.15 is without the rough motion estimation which is mentioned in Section 3.6. It is obvious that the encoder with our proposed rate control scheme outperforms JVT-G012 for buffer control. In terms of buffer overflowing and underflowing, the effect of rate control could also be observed from Table 3.6, where the scheme marked with # denotes the rough motion estimation (described in Section 3.6) being turned on. Our scheme could prevent frame skipping and bandwidth wasting more effectively. Note that due to the small buffer size for low delay applications, after intra-coding the first I frame, some P frames are always skipped to decrease the buffer occupancy below 100%. This kind of frame skipping is unavoidable for low delay communication with small size buffer. This is the reason why frame skipping also happens with our rate control scheme, albeit to a much less extent.

When frame skipping is turned on, the difference between the actual frame bits and the target frame bits in terms of MAD is compared in Table 3.6. If the current frame is skipped due to buffer overflow, its bit mismatch is not calculated in the comparison. Because if a frame is skipped, the encoder only uses a few bits to encode the “skip” information, this will induce a large mismatch between

the target frame bits and the actual encoded bits for the skipped frames, so the skipped frame is not included when calculating the MAD of frame level target bits mismatch to achieve a fair comparison. In Table 3.6 the target bits mismatch is reduced by up to 75% with our scheme.

Because of the optimized bit allocation and accurate QP adjustment, the encoder with our proposed rate control scheme could also achieve better or comparable reconstructed visual quality for most of the test sequences. Although the main advantage of our scheme is to improve the accuracy of the target bits matching, the experiment results presented in Table 3.6 shows an increase of luminance PSNR by up to 1.13dB (average around 0.33dB) at low bitrates. With a given target bitrate, if one rate control scheme skips more frames, then it would definitely allocate more bits to encode the other frames to achieve the target bitrate. In the same way, the rate control scheme with less skipped frames uses less average bits to encode its non-skipped frame, and thus achieves less PSNR value per encoded (non-skipped) frames. Using this PSNR to compare rate control schemes is unfair, because the distortion of the skipped frame should also be of concerned. In the rate control tests at MPEG-4 Visual part 2, frame repetition method is decided as the common method to calculate PSNR, because the frame is also repeatedly displayed at the decoder side if the following frame is skipped. In this chapter, this kind of PSNR calculation is used, the same as that employed in [15].

From this comparison which is carried out with various test sequences, it is found that the function of rough motion estimation (marked with # in Table 3.6 when activated) could help the encoder to perform better rate control for sequences with high motion (e.g. *FOOTBALL*), but it has not such effect on normal sequences. So it is recommended that this optional rough estimation function is turned on only when the encoding sequence has the property of continuous high

motion.

Note that *CREW* is a kind of sequence with abrupt scene changes due to camera flash, and our scheme adaptively allocates more bits to the frames with higher coding complexity. So, more bits are assigned to encode the frames with camera flash to improve the quality of such frames. This induces less bits left to encode other frames, and it is the main reason for the slight average PSNR drop. However this 0.07 dB average PSNR drop is very difficult to observe subjectively. And besides PSNR, the number of skipped frames, which is reduced by our scheme in *CREW*, is also important to evaluate subjective visual quality. So even for *CREW*, our proposed scheme also outperforms JVT-G012 with comparable PSNR value and reduced number of skipped frames.

The subjective comparison of reconstructed video quality is also shown in Figs. 3.16 and 3.17. With JVT-G012, the encoder skips the 78th and 80th frames of *FOREMAN* and the 63th and 65th frames of *FOOTBALL* due to ineffective buffer control. In contrast, our scheme provides better motion continuity.

The effect of the σ parameter, which is specified in Eqs. (3.36) and (3.37), is shown in Fig. 3.18. There is a scene change at the 131th frame of the *Table* sequence. Since human visual system always pays more attention to new scene, more bits should be allocated to the beginning of the new scene to improve its visual quality. Meanwhile, due to high quality referencing frames, the prediction of subsequent frames can be improved. As a result, the coding efficiency can be improved.

3.8 Conclusions

In this chapter, we addressed the problems associated with the existing H.264/AVC bit allocation strategy for rate control JVT-G012 [21]: inaccurate MAD prediction and inaccurate estimation of the amount of texture bits. We proposed a bit allocation scheme with adaptive MAD prediction and sum bits R-Q model for low delay application of H.264/AVC standard. Both the frame level and MB level rate control could be achieved by our scheme.

There are five new results/improvements introduced in our scheme:

- (1) A spatial MAD prediction model is presented, and it could capture the abrupt MAD changes more efficiently. The MAD could be predicted adaptively by choosing either the spatial or temporal model with the help of real-time estimation of MAD prediction. This adaptive prediction model could achieve more accurate MAD prediction at both the frame level and the MB level.
- (2) One sum bits R-Q model relating the amount of sum bits and the QP value is created by abridging the inaccurate estimation process of texture bits.
- (3) With our proposed sum bits R-Q model and the relation of QP and PSNR in the H.264/AVC standard, the QP adjustment is optimized at the MB level with Lagrange theory.
- (4) More bits are adaptively allocated to the frame when scene change or high motion happens.
- (5) Instead of using the average QP of the previous frame, the initial QP at the MB level rate control is set as the frame level predicted QP value. In this way, the MB level rate control starts from a closer range to meet the target frame bits.

From the simulation results, our proposed methods always outperforms JVT-G012 [21]. The mismatch between target frame bits and actual frame bits is reduced by up to 75%, and the buffer occupancy is much better controlled to avoid overflowing (which causes frame skipping) and underflowing (which causes bandwidth wasting). Meanwhile, both the subjective and objective reconstructed video quality are improved, and luminance PSNR is increased by up to 1.13dB (with an average of around 0.33dB) at low bitrates.

Chapter 4

Region-Of-Interest based Resource Allocation for Conversational Video Communication of H.264/AVC

4.1 Introduction

With the fast development of video communication technology, conversational multimedia services such as video telephony and video conferencing have become an important component of personal communication [13]. This kind of real-time communication requires video coding strategy with *high compression ratio* to achieve best video quality at a given bitrate, *small buffer* for low-delay interactive communication, and *low computing complexity* for easy implementation with personal video application terminals (e.g. handphone and PDA).

H.264/AVC [22] provides video compression strategy for video conferencing

and video telephony applications. The novel features of H.264/AVC significantly improve the coding efficiency when compared to the former video coding standards [22]. However, the improvement is achieved at the expense of extremely high computing complexity caused by the newly adopted technologies, especially those for motion estimation/motion compensation (ME/MC), such as rate-distortion optimization (RDO) based variable block size ME/MC, multiple reference frames ME/MC, and quarter-sample-accurate ME/MC [22]. It is difficult to integrate this type of complex encoder/decoder into personal terminals (e.g. handphone and PDA) for real-time video communication. Therefore, it is desirable to obtain a good tradeoff between the computing complexity and the compression ratio.

There are many interesting results that aim to find fast ME and mode decision algorithms to simplify the coding complexity. The motion vector field adaptive search technique (MVFAST) [73] effectively utilizes the correlation of adjacent blocks for efficient motion vector determination. The macro-block (MB) level neighborhood information is exploited in [25] along with a set of skip mode conditions. In [26], the homogeneous region detection is used to propose a fast inter-mode block-size mode decision algorithm. In [27], a fast ME algorithm is derived by excluding the low-possibility modes in the mode-decision process. The property of the all-zero coefficient block that is produced by quantization and coefficient threshold is utilized in [28] to effectively skip unnecessary modes. The scheme of [29] limits the candidate modes to a small subset by pre-encoding a down-sampled small frame. Both motion vector and mode decisions are jointly optimized in [30]. The transform domain property is utilized in [31] to achieve efficient motion mode determination. The result in [32] proposed a fast ME method based on the available motion information when transcoding H.264/AVC from MPEG2. The above fast ME and mode decision algorithms can be used to reduce

the computing complexity with slight quality loss.

Most of the fast algorithms give equal importance to every MB regardless of its relative importance to the human visual system (HVS). However, in many video applications, clients might pay more attention to the region of their interest. For example, the shoulder and head video is always encoded in real-time video communication, and the region-of-interest (ROI) of clients is usually not the background but the human face. Therefore, for both the encoder and the decoder, more resources including bits and computing power should be allocated to the human face to improve the overall visual quality. The research on the HVS has drawn a lot of attention [74–83], and the concept of ROI is an efficient tool for the classification of frame. That is, it could be used to divide a frame into several parts with different importance. When the available resources including bits and computing power are not enough, the ROI information can be used to optimally allocate the available resources to different parts of the frame according to their relative importance. In this way, the overall visual quality could be optimized.

There are three major problems when the concept of ROI is used in video coding: the detection of ROI, the ROI based bits allocation, and the ROI based computing power allocation. The former two have been well addressed and there are many interesting results. In [84] and [85], ROI is detected according to a two-level neural network classifier. The method of [86] utilizes the method of fuzzy logic control to adaptively assign the weighting factors to each MB, hence different bits are allocated to each MB according to its complexity of the rate distortion model. Most of the existing schemes [84–87] detect the ROI after ME, and the coding parameter (e.g. quantization parameter) is next adjusted to utilize more bits to encode the ROI. These existing ROI detection schemes cannot be directly used in H.264/AVC because there is a dilemma in the detection of ROI due to

the employment of RDO in the H.264/AVC. Motion information is very important for the detection of ROI and is available after the RDO. The RDO can only be performed after the value of QP is available. However, the value of QP can be calculated after the ROI is detected. A similar dilemma also exists in rate control of H.264/AVC [21], and it is properly solved by using the available spatial and temporal MAD to estimate the actual MAD in our previous work [24], and as presented in Chapter 3. Meanwhile, automatic ROI extraction based on the HVS is a rather difficult task. In the recent research of human visual sensitivity [87], color contrast, texture contrast, motion, skin color and face detection are separately exploited and integrated together to reflect the processing ability of HVS. But this ROI detection algorithm is too complex to be implemented into real-time video communication. Moreover, the ROI based computing power allocation problem has not been well studied yet. It is necessary to consider it because the computing capacity of handheld devices is very limited.

We address the above three problems in this chapter. A fast ROI detection scheme is proposed to detect the ROI before the ME, and the dilemma in the detection of ROI is thus solved. The direct frame difference and the skin-tone information are applied to detect the ROI. This detection scheme is very suitable for conversational video communication due to its simple derivation. An ROI based rate control scheme is then designed for H.264/AVC and this scheme is an improved version of that presented in Chapter 3. In this new scheme, a relatively larger portion of bits is assigned to encode the ROI such that the overall visual quality could be improved. Since this chapter focuses on the conversational video communication, and the computing power of handheld device is usually limited, a simplified computing power allocation scheme is provided with the concern of both encoding and decoding complexity. Specifically, several coding parameters

(e.g. ME modes, sub-pixel ME accuracy, ME search range and multiple reference frames ME) should also be properly adjusted to allocate more computing power to the ROI. The overall encoding time is saved due to the reduction of computing complexity in the region outside the ROI.

The remainder of this chapter is organized as follows. Section 4.2 presents our ROI detection scheme for H.264/AVC standard. In Section 4.3, the ROI based bit allocation is implemented with the rate control scheme proposed in Chapter 3. Computing power allocation with the determined ROI is described in Section 4.4. With the proposed ROI detection method and the parameter adjustment strategy, our simplified system is compared to the JVT H.264/AVC reference software JM9.8, where the experimental results and discussions are shown in Section 4.5. Finally, the concluding remarks are given in Section 4.6.

4.2 Fast ROI Detection

4.2.1 Dilemma in the Detection of ROI

The psychology of HVS has been studied for more than 100 years [88]. The achievements on visual attention can be used to process the video signal adaptively by allocating more noise or distortion to the frequency or spatial area which is less noticeable to the HVS, and improve the visual quality to the frequency or spatial area in which the HVS is more interested in. In the recent research of visual sensitivity [87], a HVS model is set up by using the features of color contrast, texture contrast, motion information, skin color and face detection. Although the integration of these features could properly indicate the perceptual quality significant map (PQSM) of HVS, the algorithms to extract these features [89–93] are extremely complex and are not suitable for implementation with the real-time video

encoder with limited computing power.

Moreover, due to the RDO in the H.264/AVC, there is a dilemma in the detection of ROI. Motion information which is the most important feature for the detection of ROI cannot be determined without the value of QP. However, the value of QP would be further adjusted after the ROI is detected.

Although similar dilemma in the rate control scheme of H.264/AVC has been well known [21], the dilemma in the detection of ROI in H.264/AVC has not been addressed yet.

4.2.2 Simple ROI Detection Method for Conversational Video Communication

In our scheme, the ROI information is derived from pixel difference and skin-tone indicator.

To solve the dilemma in the detection of ROI, the difference between consecutive frames before ME can be applied to design a fast ROI detection scheme. As shown in Figs. 4.1(b), 4.2(b) and 4.3(b), the temporal difference of two consecutive frames could be used to approximately indicate the texture gradient and motion extension which are important when creating PQSM [87].

Meanwhile, for the conversational application, skin-tone area should also be highlighted as the HVS always pay more attention to the skin area (e.g face) even without motion. In the video compression standards, the YC_bC_r color space is widely used [10,11,17,20,22]. Fortunately, the YC_bC_r color space is similar to the tint-saturation-luma (TSL) color space [94] which provides the best face detection result as compared to other color spaces (e.g. red-green space [95]), and it is perceptually uniform [96]. Although some studies assumed that the chrominance parts of the skin-tone are independent of the luminance part [97–99], in fact the

4.2 Fast ROI Detection

89

skin-tone color is nonlinearly dependent on luminance in practice [100]. In our scheme, the skin-tone detection is based on the model proposed in [100], where the skin-tone color is nonlinearly dependent on luminance. The skin-tone model is presented by the centers ($\bar{U}(Y)$ and $\bar{V}(Y)$) and spread of the cluster ($W_U(Y)$ and $W_V(Y)$). The transformed chrominance could be calculated as:

$$V'(Y) = \begin{cases} V(Y) & : Y \in [125, 188] \\ (V(Y) - \bar{V}(Y)) \cdot \frac{38.76}{W_V(Y)} + 154 & : \text{else} \end{cases} \quad (4.1)$$

$$U'(Y) = \begin{cases} U(Y) & : Y \in [125, 188] \\ (U(Y) - \bar{U}(Y)) \cdot \frac{46.97}{W_U(Y)} + 108 & : \text{else} \end{cases} \quad (4.2)$$

where Y is the luminance value of the pixel, $U(Y), V(Y)$ are the corresponding chrominance values, $U'(Y), V'(Y)$ are the nonlinear functions between luminance and chrominance values, and

$$W_V(Y) = \begin{cases} 0.172V(Y) + 17.25 & : Y < 125 \\ -0.612V(Y) + 153.8 & : Y > 188 \end{cases} \quad (4.3)$$

$$W_U(Y) = \begin{cases} 0.22U(Y) + 19.48 & : Y < 125 \\ -0.701U(Y) + 178.85 & : Y > 188 \end{cases} \quad (4.4)$$

$$\bar{U}(Y) = \begin{cases} -0.09U(Y) + 119.47 & : Y < 125 \\ 0.213U(Y) + 68 & : Y > 188 \end{cases} \quad (4.5)$$

$$\bar{V}(Y) = \begin{cases} -0.09V(Y) + 165.47 & : Y < 125 \\ 0.213V(Y) + 114 & : Y > 188 \end{cases} \quad (4.6)$$

The skin-tone pixel is detected if the following elliptical model is satisfied for the transformed chrominance:

$$\frac{(x - m_x)^2}{p_x^2} + \frac{(y - m_y)^2}{p_y^2} \geq 1 \quad (4.7)$$

where

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos \alpha & \sin \alpha \\ -\sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} U'(Y) - q_x \\ V'(Y) - q_y \end{bmatrix} \quad (4.8)$$

and $m_x = 1.6, m_y = 2.41, p_x = 25.39, p_y = 14.03, \alpha = 2.53 \text{ rad}, q_x = 109.38$ and $q_y = 157.02$. These parameters are set according to the training result from HHI image database. The details of this part could be found in [100]. The detected skin-tone areas are shown in Figs. 4.1(c), 4.2(c) and 4.3(c). Because some area on the face (i.e. eyes, mouth) may not be detected as skin, one low pass filter is further employed to dilate the skin-tone area to cover a larger face area, the dilated skin-tone areas are given in Figs. 4.1(d), 4.2(d) and 4.3(d). Note that because HVS always focuses on the center of the frame [101], the MBs on the boundary of frame is not included into the skin-tone detection.

Because the video sequence is encoded at the MB level, the pixel level direct difference and skin-tone information are combined to indicate the ROI importance at the MB level. Let p be a binary parameter to represent whether the pixel is

Figure 4.1: Example 1 of ROI detection: (a) the 5th original frame of the FOREMAN sequence at QCIF spatial resolution; (b) temporal difference at pixel level; (c) the detected skin-tone areas; (d) the dilated skin-tone area; (e) I_{ROI} at MB level; (f) initial detected ROI mask; (g) final detected ROI mask after mean filtering.

detected as skin-tone (1:= yes, 0:=no), then

$$I_{ROI}[i] = \sum_{j=1}^{16} \sum_{k=1}^{16} [|F0[i, j, k] - F1[i, j, k]| \cdot (p[i, j, k] + 0.5)] \quad (4.9)$$

where i is the MB index, j, k are the pixel index within one MB, $F0$ and $F1$ are two continuous frames, and I_{ROI} is the ROI indicator at MB level. In Figs. 4.1(e), 4.2(e) and 4.3(e), I_{ROI} of each MB in the testing frame is described in a grey-level graphic, where the white color denotes the maximal I_{ROI} of the testing frame and the black one denotes the minimal value.

The procedure for creating the ROI mask with the I_{ROI} information is given in the following three steps:

Figure 4.2: Example 2 of ROI detection: (a) the 22th original frame of the MISS sequence at QCIF spatial resolution; (b) temporal difference at pixel level; (c) the detected skin-tone areas; (d) the dilated skin-tone area; (e) I_{ROI} at MB level; (f) initial detected ROI mask; (g) final detected ROI mask after mean filtering.

(1) The I_{ROI} at MB level is first scaled as

$$\mu[i] = \frac{I_{ROI}[i]}{\bar{I}_{ROI}} \quad (4.10)$$

where i is the MB index of the current frame, $\mu[i]$ represents the importance of the i -th MB to detect ROI, $\bar{I}_{ROI}$ is the average I_{ROI} of all the MBs in the current frame. $\mu[i]$ is next bounded as

$$\mu[i] = \max\{\mu_L, \min\{\mu[i], \mu_U\}\} \quad (4.11)$$

where μ_L and μ_U are two constants to denote the lower and upper boundaries of μ , set to 1.0 and 2.0 respectively in our scheme. $\mu[i]$ is further re-mapped

Figure 4.3: Example 3 of ROI detection: (a) the 73th original frame of the M&D sequence at QCIF spatial resolution; (b) temporal difference at pixel level; (c) the detected skin-tone areas; (d) the dilated skin-tone area; (e) I_{ROI} at MB level; (f) initial detected ROI mask; (g) final detected ROI mask after mean filtering.

to the range $\{-\frac{\mu_U - \mu_L}{2}, \frac{\mu_U - \mu_L}{2}\}$ as

$$\mu[i] = \mu[i] - \frac{1}{2}(\mu_U + \mu_L) \quad (4.12)$$

- (2) With the assumption proposed in [101] that HVS always focuses on the center of the frame, a frame boundary ROI mask is introduced as shown in Fig. 4.4. For the sequences with QCIF spatial resolution, the MBs in the upper row, and left and right columns are excluded outside the ROI region. For the CIF spatial resolution sequences, the boundaries are extended with one more MB width. The μ values of the MBs within the boundaries (MBs with black color in Fig. 4.4) are set to $-\frac{\mu_U - \mu_L}{2}$.
- (3) The initial ROI masks for different test sequences after the above two steps

Figure 4.4: The boundary ROI mask of one frame.

are shown as grey-level graphic in Figs. 4.1(f), 4.2(f) and 4.3(f), where white and black colors respectively denote the maximal and minimal μ values of each MB. Although these initial ROI masks could indicate the ROI region well, the μ values of the spatial neighboring MBs within a local area sometimes varies significantly. If the encoder allocates the coding resources to the ROI according to the μ values within these masks, the visual quality of the MBs with high μ values might be a lot better than their neighbors. This will induce block artifacts at the edge of ROI and adversely affect the overall visual quality of the whole frame. To solve this problem, a weighted 3×3 mean filter M is applied to all the MBs inside the boundary (MBs with white color in Fig. 4.4) to smoothen the initial ROI mask:

$$M = \begin{bmatrix} \frac{1}{12} & \frac{1}{12} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{3} & \frac{1}{12} \\ \frac{1}{12} & \frac{1}{12} & \frac{1}{12} \end{bmatrix} \quad (4.13)$$

The final detected ROI masks for different test sequences after the above three steps are shown in Figs. 4.1(g), 4.2(g) and 4.3(g), where the μ value of each MB is also specified into grey-level from white (maximal value of μ , i.e. $\frac{\mu_U - \mu_L}{2}$) to black (minimal value of μ , i.e. $-\frac{\mu_U - \mu_L}{2}$). The block artifact is effectively

removed by the weighted mean filter.

The above proposed scheme only utilizes the direct frame difference and the skin-tone information, so it could be easily implemented into an H.264/AVC encoder without the involvement of QP to realize fast ROI detection for real-time video communication. Meanwhile, this ROI detection scheme could achieve impressive ROI mask of the HVS to prepare for further coding parameter adjustments.

Note that our ROI detection scheme is very effective for the conversational video, where the background is usually still and clients are more interested in the human face of the foreground. When the proposed ROI detection cannot work effectively for some other kinds of videos, equal importance could be given to each MB, or higher importance to the center region of the frame as the HVS always focuses on the center of the frame [101].

4.3 ROI Based Bit Allocation for H.264/AVC

Bit allocation for rate control is one of the most important open issues in video encoding. In Chapter 3, we have proposed a bit allocation scheme for H.264/AVC rate control, but this scheme allocate bits with equal importance to every MB of the frame, regardless of its relative importance to the HVS. In fact, to encode the ROI with higher quality, more bits should be allocated to the MB which is more important to the HVS.

Based on the classic rate-distortion theory [66], RDO is employed in the H.264/AVC standard [23] to provide optimal bit allocation between motion information and residual information. An encoder with a smaller Lagrangian parameter usually could generate more encoding bits and achieve better visual quality. So smaller

QP value should be assigned to MBs in the ROI. The bit allocation with ROI concern could be easily implemented with our previously proposed rate control scheme for H.264/AVC, as described in Chapter 3.

To provide an ROI based rate control at the MB level, Eq. (3.27) is replaced by

$$\max \sum_{i=1}^N \{(a \times Q_p[i] + b)w[i]\} \quad (4.14)$$

where $w[i]$ is the weighting factor of the i -th MBs. It can be chosen to incorporate the importance of the i -th MB's distortion as

$$w[i] = (2\mu[i] + \mu_U)^2 \quad (4.15)$$

Therefore, a larger $w[i]$ is selected if the i -th MB is in the ROI. Similar to Eq. (3.45), the value of $Q_p[i]$ is derived as

$$\begin{aligned} Q_p[i] &= \frac{X1[i]}{T[i] - (N - i)X2[i]} \sqrt{MAD_{pred,adapt}[i]} \\ &\quad * \sum_{k=i}^N \sqrt{\frac{w[k]}{w[i]} MAD_{pred,adapt}[k]} \end{aligned} \quad (4.16)$$

4.4 ROI Based Computing Power Allocation

The H.264/AVC [42] [22] is the latest international video coding standard which could achieve a significant improvement in the rate-distortion efficiency. The H.264/AVC could provide half bit-rate savings compared with MPEG-2 [10] which is the most common standard used for video storage and transmission, and the compression gain of H.264/AVC over H.263 [11] is in the range of 25% to 50% due to different type of applications.

When compared with the previous video coding standards, many new methods [42] [22] are introduced in the H.264/AVC to enhance its compression performance, such as *variable block size motion estimation*, *quarter-sample-accurate motion estimation* and *multiple reference frames motion estimation*. The complexity of H.264/AVC encoder and decoder are significantly increased due to the employment of these new methods. Moreover, the personal video terminals (e.g. handphone and PDA) of the clients are always equipped with relatively low computing capacity. Clearly, it is very difficult to use the original H.264/AVC encoder and decoder for real-time conversational video communication. Therefore, both the encoder and the decoder should be properly simplified by adaptively adjusting the coding parameters with the ROI concern. These are discussed below.

4.4.1 Encoder complexity

4.4.1.1 MB candidate modes

To represent more accurate motion field of moving objects to reduce the residual error, there are a total of 7 different MB partitions (16×16 , 16×8 , 8×16 , 8×8 , 8×4 , 4×8 , and 4×4) that are utilized in the H.264/AVC [42] [22], as shown in Fig. 4.5 [102].

For the MBs in inter frame, both the inter prediction and intra prediction are tried to select the prediction mode with minimal R-D (rate distortion) cost.

For inter prediction, each MB could be first split up in four ways, as shown in Fig. 4.5 (a): either one 16×16 MB partition, two 16×8 partitions, two 8×16 partitions or four 8×8 partitions. If the 8×8 mode is chosen from these four modes with minimal R-D cost, each of the four 8×8 sub-MB within the MB could be further split in additional four ways, as shown in Fig. 4.5 (b): either as one 8×8 sub-MB partition, two 8×4 sub-MB partitions, two 4×8 sub-MB partitions

4.4 ROI Based Computing Power Allocation

98

Figure 4.5: The MB partitions of H.264/AVC.

4.4 ROI Based Computing Power Allocation

99

or four 4×4 sub-MB partitions. These MB partitions and sub-MB partitions provide a large number of possible combinations within each MB. This kind of tree structured motion compensation partitions MB into motion compensated sub-blocks of variable size.

Figure 4.6: The intra prediction methods for 4×4 partition of H.264/AVC .

For each MB, the intra prediction is carried out for each 4×4 block or for the whole 16×16 macro-block. There are a total of nine intra prediction modes for each 4×4 block, as shown in Fig. 4.6 [102]:

- Mode 0: Vertical prediction
- Mode 1: Horizontal prediction
- Mode 2: DC prediction
- Mode 3: Diagonal down-left prediction
- Mode 4: Diagonal down-right prediction
- Mode 5: Vertical-right prediction
- Mode 6: Horizontal-down prediction
- Mode 7: Vertical-left prediction

4.4 ROI Based Computing Power Allocation

100

- Mode 8: Horizontal-up prediction

Figure 4.7: The intra prediction methods for 16×16 partition of H.264/AVC .

And there are four modes for a 16×16 macro-block, as shown in Fig. 4.7 [102]:

- Mode 0: Vertical prediction
- Mode 1: Horizontal prediction
- Mode 2: DC prediction
- Mode 3: Plane prediction based on a linear spatial interpolation.

The intra prediction is spatially based on previously encoded and reconstructed blocks, and the residual information is subtracted from the current block prior to encoding.

Therefore, H.264/AVC provides more flexible prediction modes for each MB in the inter frames from the set {INTRA16x16, INTRA4x4, INTER16x16, INTER16x8, INTER8x16, INTER8x8, INTER8x4, INTER4x8, INTER4x4, SKIP, DIRECT}. Both the motion vector and the MB partition mode should be determined with RDO using the Lagrangian method, and the final MB mode is determined with the minimal R-D cost. However, RDO based ME and mode decision require a series of operations (e.g. motion vector estimation, quantization, integer transform, entropy coding, inverse quantization and inverse integer transform) for

each prediction mode. Because of the many possible partitions of each MB, it is extremely computationally intensive to select the best mode from all the candidate modes. Subsequently, the resultant computational burden challenges the personal terminals (e.g. handphone and PDA) for real-time video communication. Due to the limited computing power of the encoder, more coding resources should be utilized on the coding of MBs in the ROI to guarantee the overall visual quality. Therefore, it is desired to design an ROI based computing power allocation scheme.

We divide all the possible MB partition modes into three subsets: subset 0 is {INTER16x16, SKIP, DIRECT}, subset 1 is { INTER16x8, INTER8x16, INTER8x8} and subset 2 is {INTRA16x16, INTRA4x4, INTER8x4, INTER4x8, INTER4x4}. As shown in Fig. 4.8 and Fig. 4.9, the partition subset with complicated modes takes more percentage in the area with higher μ values. Based on this observation, a simple algorithm is used to predefine the candidates for mode decision as follows:

if $\mu[i] = -\frac{\mu_U - \mu_L}{2}$, **then**

The partition mode is chosen from subset 0,

else

if $-0.5 < \mu[i] < -0.25$, **then**

The partition mode is selected from subsets 0 and 1,

else

The partition mode is chosen from subsets 0, 1 and 2.

end if

end if

In this way, most of the available computing power is allocated to the ROI and the encoder is simplified by reducing the encoding complexity of MBs in the region

4.4 ROI Based Computing Power Allocation

102

Figure 4.8: The original testing frames are shown in the left column and the final detected ROI mask and the MB partition modes for the same frames are in the right column. Random analysis of (a) the 8th frame of FOREMAN@QCIF 15Hz, (b) the 17th frame of MISS@QCIF 30Hz, (c) the 16th frame of M&D@QCIF 30Hz.

Figure 4.9: The percentage of the partition subsets for the MBs within the same range of μ value. Analysis of FOREMAN@QCIF 15Hz with QP=40.

outside the ROI. This scheme would induce slight difference between the derived MB mode and the actual best mode through a full search for the MBs in the region outside the ROI. However, this would not adversely affect the overall subjective visual quality.

4.4.1.2 The number of reference frame

Figure 4.10: The method of multiple reference frames for H.264/AVC.

Contrary to the previous standards which only allows the use of the frame just before and/or after the current frame as the referencing one(s), the H.264/AVC standard enhances the coding efficiency by offering the option of multiple reference frames in the inter-frame coding. As shown in Fig. 4.10 [22], besides the motion vector, the frame reference parameters (Δ) are also transmitted to the decoder side. This concept could also be extended to bi-directional predicted frames. By analyzing the reference frames of sequence FOREMAN@QCIF 15Hz with QP=40, the other reference frames (not the one just before and/or after the current frame) only occupy 4.3% of the non-ROI region ($\mu[i] = -\frac{\mu_U - \mu_L}{2}$), and increased to 25.9% of the ROI ($\mu[i] > -\frac{\mu_U - \mu_L}{2}$) area. Therefore, to save the coding complexity of MBs in the region outside the ROI, the reference frames of the MBs in the non-ROI region are fixed to the frame just before and/or after the current frame. The MBs in the ROI could use multiple previously and/or subsequently coded frames as referencing ones to maximize their qualities. The number of the candidate reference frames could be pre-defined before encoding.

4.4.1.3 Sub-pixel accurate motion estimation

The H.264/AVC standard uses quarter-sample-accurate motion vectors while most of the previous standards only enable at most half-sample-accurate motion vectors. The more accurate motion prediction provides the encoder a significant improvement of motion estimation and motion compensation. The pixel value at the sub-pixel position needs to be interpolated as the method illustrated in Fig. 4.11 [22], in which the original integer pixel is labelled by upper-case letters “A-U” and the interpolated sub-pixel are represented by other symbols. The half-sample-accurate pixels (e.g. b , h , m , s , cc , dd , ee , ff) are interpolated from the integer pixel with a one-dimensional six-tap FIR filter. For example, the value at the location b in

Figure 4.11: The method of interpolation for H.264/AVC.

4.4 ROI Based Computing Power Allocation

106

Fig. 4.11 is first calculated with the values at the integer locations E, F, G, H, I and J :

$$b = (E - 5F + 20G + 20H - 5I + J) \quad (4.17)$$

The final interpolation value is further obtained as follows:

$$b = (b + 16) >> 5 \quad (4.18)$$

and

$$b = \min(255, \max(b, 0)) \quad (4.19)$$

The half-sample-accurate pixel value at position j could be achieved by using the one-dimensional six-tap FIR filter on the six samples at half-sample-accurate position as

$$j = (cc - 5dd + 20h + 20m - 5ee + ff) \quad (4.20)$$

$$j = (j + 16) >> 5 \quad (4.21)$$

$$j = \min(255, \max(j, 0)) \quad (4.22)$$

In this way, to interpolate j , the values at the half-sample-accurate position (cc, dd, h, m, ee, ff) must be calculated first. Compared to other calculations for the half-sample-accurate pixel, the interpolation for j is the most complex.

The values at the quarter-sample-accurate position, denoted as a, c, d, n, f, i, k , and q , are calculated by averaging the two nearest samples at integer and half

4.4 ROI Based Computing Power Allocation

107

sample positions, such as

$$a = (G + b + 1) \gg 1 \quad (4.23)$$

In a similar way, the values at quarter-sample-accurate position denoted as e , g , p , and r are derived by averaging the two nearest samples at half sample positions in the diagonal direction as

$$e = (b + h + 1) \gg 1 \quad (4.24)$$

Table 4.1: The Calculation Complexity at Different Sub-pixel Position.

Example position	sub-pixel accuracy	interpolation complexity
H	integer	no
b, h	half	$1 \times (\text{six-tap})$
j	half	$7 \times (\text{six-tap})$
a, c, d, n	quarter	$1 \times (\text{six-tap}) + 1 \times (\text{two-tap})$
e, g, p, r	quarter	$2 \times (\text{six-tap}) + 1 \times (\text{two-tap})$
i, f, k, q	quarter	$7 \times (\text{six-tap}) + 1 \times (\text{two-tap})$

From the above derivations, it is clear that the sub-pixel values at different positions are with different interpolation complexity. The calculation complexity at different sub-pixel positions are given in Table 4.1.

By analyzing the sub-pixel ME of sequence MISS@QCIF 15Hz with QP=40, the sub-pixel (both half- and quarter-pixel) ME only occupies 2.1% of the non-ROI region, and increased to 42.7% of the ROI area. To achieve fast video coding for real-time communication, the accuracy of motion vectors for the MBs in the non-ROI region is up to half-pixel, and the MBs in the ROI employ the quarter-pixel motion estimation. Note that the sub-pixel ME in the ROI area could be further optimized with the concern of decoding interpolation complexity, as described later

in Section 4.4.2.

4.4.1.4 Motion estimation search range

Figure 4.12: The relation of the percentage of motion vectors belongs to large search range and the μ value. Analysis of FOREMAN@QCIF 15Hz with QP=40.

Let $S[i]$ be the search range of the i -th MB, and $\tilde{S}$ be a predefined constant that specifies the search range. As shown in Fig. 4.12, the maximal motion vectors belonging to a larger search range ($> 0.5\tilde{S}$) occupies a higher percentage in area with higher μ values. Therefore, the search range of MB i is adjusted by

$$S[i] = \tilde{S} * (0.75\mu[i] + 0.625) \quad (4.25)$$

Obviously, a larger range could be searched for the MB with higher importance to HVS, and the coding complexity of the less important region is saved.

4.4.2 Decoder complexity

The analysis of computing complexity of H.264/AVC decoder is presented in [103], the interpolation caused by sub-pixel ME is the major time consuming part of the

decoder, it could take up to 56% of the decoding time. In the conversational video communication, the computing power of user's terminals are limited. Besides the predefined sub-pixel ME constraints presented in Section 4.4.1.3, it is desired to proposed an ROI based decoding-friendly scheme for the ROI region ($\mu > -0.5$), so as to allocate more decoder resources on the area with high ROI interest (high μ value). Therefore, the decoding complexity should be optimized at the encoder side.

In H.264/AVC, rate distortion optimization (RDO) is used to determine motion information [23], both motion estimation and mode decision should be optimized by using the Lagrangian method, as the description of Eqs. (3.9, 3.10, 3.11, 3.12) in Section 3.2.

Inspired by the previous computing power allocation work proposed in [104], we take the decoding complexity into account by modifying the above rate-distortion (R-D) cost functions of Eqs. (3.9) and (3.10) into rate-distortion-complexity (R-D-C) cost functions:

$$J_{mv} = \{D_{mv} + \lambda_{motion} \cdot R_{mv} + \beta_{motion} \cdot C_{motion}\} \quad (4.26)$$

$$J_{mode} = \{D_{rec} + \lambda_{mode} \cdot (R_{mv} + R_{coeff}) + \beta_{mode} \cdot C_{mode}\} \quad (4.27)$$

where β_{motion} and β_{mode} are the two Lagrange multipliers to compromise the decoder complexity during motion vector and motion mode determination respectively, C_{motion} and C_{mode} are the parameters to indicate the decoding complexity at motion vector and motion mode determination stages respectively. Here, C_{motion} is calculated based on the interpolation complexity of the current motion vector, and C_{mode} is the sum of the interpolation complexity of all the motion vectors in

the current mode.

As described in Section 4.4.1, it is known that the sub-pixel values at different positions are with different interpolation complexity. Based on the details in Table 4.1, the interpolation complexity of each motion vector C_{motion} is predefined as

$$C_{mode}(MV) = \begin{cases} 0 & : \text{MV is at integer position} \\ 1 & : \text{MV is at position b, h} \\ 1.5 & : \text{MV is at position a, c, d, n} \\ 2.5 & : \text{MV is at position e, g, p, r} \\ 7 & : \text{MV is at position j} \\ 7.5 & : \text{MV is at position i, f, k, q} \end{cases} \quad (4.28)$$

The higher values of β_{motion} and β_{mode} could save more decoder complexity at the cost of video quality degradation. To save more decoder resources from the ROI according to each MB's relative importance, β_{motion} and β_{mode} are adjusted as follows:

$$\beta_{motion}[i] = 80 \cdot (0.5 - \mu[i]) \quad (4.29)$$

$$\beta_{mode}[i] = 4 \cdot (0.5 - \mu[i]) \quad (4.30)$$

In this way, the areas with lower μ values could be decoded with less interpolation complexity.

4.5 Experimental Results

The performance of ROI based resource allocation scheme is evaluated in this section. Several standard MPEG test sequences with QCIF (176×144) spatial resolution are carried out in our experiments. The simulation is implemented with the latest JVT reference H.264/AVC software JM9.8 [62], in which JVT-G012 [21] is adopted to regulate the bitrates while the rate control scheme proposed in Section 4.3 is employed in our encoder. Since real-time conversational video communication is studied in this chapter, all the test sequences are intra-coded for the first frame (I frame) and followed with inter-coded frames (P frames). The buffer size is set as $0.1 * \text{bitrate}$, i.e. the maximal buffer delay is limited to 100ms to satisfy the conversational low-delay requirement. One additional buffer size is also simulated to increase the buffer delay to 500ms. Other coding parameters are listed as follows:

- Motion estimation search range is set to 16 pixels.
- RDO is enabled.
- CABAC is enabled.
- The number of reference frames is set to 2.
- Quarter-sample-accurate motion estimation is enabled.

Note that some of the above coding parameters should be adaptively adjusted for our ROI based computing power allocation scheme.

When the buffer overflowed, the encoder skips frames until there is available space in the buffer. When the buffer underflowed, certain bandwidth will be wasted. Thus, it is desired to design a good buffer control scheme to reduce and even prevent the buffer from overflowing and underflowing. Clearly, it can be

4.5 Experimental Results

112

Table 4.2: Comparison of Our Scheme and JM9.8 with Coding Complexity Optimization.

sequence	Sch.	Tgt. bitr. (kb/s)	Actl. bitr. (kb/s)	Frame bits Mism.	ME time (ms/f)	overall time (s/f)	overall PSNR (dB)	ROI PSNR (dB)	IP comp. $\times 10^3$
FOREMAN QCIF, 15Hz	JM98	19	19.05	484	1006	2.27	27.53	26.75	495
	Ours		19.04	188	271	0.91	27.10	27.02	253
	JM98	64	64.03	1168	961	2.85	34.18	33.36	463
	Ours		64.03	492	457	1.97	33.86	33.67	312
	JM98	244	244.01	1358	927	4.16	41.49	40.59	474
	Ours		244.02	1134	574	2.77	41.07	40.86	344
MISS QCIF, 30Hz	JM98	8	8.04	155	978	2.17	32.83	30.13	59
	Ours		8.04	105	388	0.96	32.82	30.39	51
	JM98	11	11.01	840	965	2.08	34.75	31.84	78
	Ours		11.04	139	404	0.88	34.61	32.15	73
	JM98	68	68.20	1308	930	2.50	42.49	40.04	148
	Ours		68.03	259	376	1.32	42.23	40.16	120
M&D QCIF, 30Hz	JM98	11	11.04	228	951	2.13	28.98	27.13	59
	Ours		11.05	162	326	0.89	28.74	27.53	48
	JM98	22	22.09	1348	938	2.32	31.18	28.64	98
	Ours		22.05	372	368	0.87	30.85	28.88	72
	JM98	118	118.20	2156	928	3.23	38.47	36.19	152
	Ours		118.07	594	342	1.54	37.98	36.49	123

shown from Table 4.2 that this is achieved by our ROI based rate control scheme. The mismatch between the actual frame bits and the target frame bits is reduced by up to 80% with our proposed encoder, and our ROI based rate control scheme could also be used to match the target bitrates as exactly as JM9.8. The time used to achieve ME/MC per frame is also specified in Table 4.2. Clearly, our proposed encoder could save the ME/MC time by up to 73%, save the overall coding time by up to 63%, and it makes the adaptive encoder more suitable for real-time application with limited computing power. Meanwhile, as the decoding interpolation complexity is also considered with ROI importance at the encoder side, the interpolation complexity is reduced by up to 48% with the R-D-C model described in Section 4.4.2.

Although the average PSNR values of the *M&D* and *FOREMAN* sequences have dropped slightly, the visual quality has improved significantly. This implies that the PSNR value sometimes cannot indicate the subjective visual quality especially at low bitrates, because some annoying features such as block and ringing

artifacts cannot be indicated by the PSNR value. The average PSNR value is calculated with the same priority for all MBs without considering the sensitivity of HVS. According to the previous research results of the perception psychopys [74] [75] [76], the difference of perceptive sensitivity to the ROI and the region outside the ROI may vary from 0 to 9.4 dB. Therefore, the use of the average PSNR value to evaluate the reconstructed visual quality is not sufficient. We also compare the PSNR value of the ROI region. Our scheme with ROI concern could improve the PSNR value of the ROI region by up to 0.40dB.

Three examples of subjective visual quality comparisons are shown in Figs. 4.13, 4.14 and 4.15. It could be observed from these figures that the detected ROI masks could represent the sensitive region of HVS very well. Although in the first and third examples the PSNR value of the reconstructed frames with our simplified encoder are slightly lower than those resulted by JM9.8, the subjective qualities are obviously improved. Note that, we have selected some frames with slight PSNR loss to indicate that even with a lower PSNR value the proposed adaptive encoder could achieve better visual quality than JM9.8. With further detailed comparison of the background, some slight distortion within the region outside the ROI could be observed, such as the edge of the collar and the cap in Fig. 4.13 and the edge of the coach in Fig. 4.15. Fortunately this noise is not sensitive to HVS and would not annoy the video clients.

Finally, to test our scheme without the coding complexity optimization, the encoder with only the ROI based rate control is simulated, and the results are given in Table 4.3. Compared with JM98, the proposed ROI based rate control without coding complexity adjustment could reduce the target frame bits mismatch by up to 83%, and improve the PSNR of ROI region by up to 0.98dB. Meanwhile, the motion estimation time and decoding interpolation complexity remains comparably

4.5 Experimental Results

114

Figure 4.13: Example 1 of comparison of the subjective visual quality. Analysis of the 9th frame of FOREMAN@QCIF 15Hz, (a) original frame, (b) detected ROI mask, (c) the reconstructed frame without ROI concern with $\text{PSNR} = 27.30$ dB, (d) the reconstructed frame with ROI concern with $\text{PSNR} = 26.72$ dB, (e) details of the most sensitive region of (c), (f) details of the most sensitive region of (d).

4.5 Experimental Results

115

Figure 4.14: Example 2 of comparison of the subjective visual quality. Analysis of the 30th frame of MISS@QCIF 30Hz, (a) original frame, (b) detected ROI mask, (c) the reconstructed frame without ROI concern with PSNR = 34.04 dB, (d) the reconstructed frame with ROI concern with PSNR = 34.19 dB, (e) details of the most sensitive region of (c), (f) details of the most sensitive region of (d).

4.5 Experimental Results

116

Figure 4.15: Example 3 of comparison of the subjective visual quality. Analysis of the 64th frame of M&D@QCIF 30Hz, (a) original frame, (b) detected ROI mask, (c) the reconstructed frame without ROI concern with PSNR = 27.26 dB, (d) the reconstructed frame with ROI concern with PSNR = 26.61 dB, (e) details of the most sensitive region of (c), (f) details of the most sensitive region of (d).

Table 4.3: Comparison of JM9.8 and Our Scheme without Coding Complexity Optimization.

sequence	Sch.	Tgt. bitr. (kb/s)	Actl. bitr. (kb/s)	Frame bits Mism.	ME time (ms/f)	overall time (s/f)	overall PSNR (dB)	ROI PSNR (dB)	IP comp. $\times 10^3$
FOREMAN QCIF, 15Hz	JM98	19	19.05	484	1006	2.27	27.53	26.75	495
	Ours		19.02	191	977	2.08	27.79	27.48	392
	JM98	64	64.03	1168	961	2.85	34.18	33.36	463
	Ours		64.03	487	939	2.70	34.04	34.07	453
	JM98	244	244.01	1358	927	4.16	41.49	40.59	474
	Ours		244.02	1096	914	4.00	41.27	41.57	471
MISS QCIF, 30Hz	JM98	8	8.04	155	978	2.17	32.83	30.13	59
	Ours		8.05	106	950	1.75	32.81	30.50	53
	JM98	11	11.01	840	965	2.08	34.75	31.84	78
	Ours		11.05	140	917	1.51	34.73	32.17	74
	JM98	68	68.20	1308	930	2.50	42.49	40.04	148
	Ours		68.07	209	917	2.27	42.31	40.16	147
M&D QCIF, 30Hz	JM98	11	11.04	228	951	2.13	28.98	27.13	59
	Ours		11.04	165	918	1.72	28.75	27.58	57
	JM98	22	22.09	1348	938	2.32	31.18	28.64	98
	Ours		22.05	352	890	1.59	30.91	28.97	86
	JM98	118	118.20	2156	928	3.23	38.47	36.19	152
	Ours		118.06	564	877	2.78	38.24	36.77	136

the same. The performance of the proposed rate control scheme without ROI concern has already been well discussed in Chapter 3.

The simulation results show that our proposed encoder with ROI concern could adaptively utilize the limited computing power to achieve better visual quality with shorter processing time than JM9.8. Our simplified encoder is thus more suitable for real-time conversational communication.

4.6 Conclusions

In this chapter, we have proposed an ROI based H.264/AVC system for real-time conversational video communication. It can be applied to meet desirable requirements such as high compression ratio, low computing complexity and low-delay transmission. In our scheme, the ROI mask is first detected with direct frame difference and skin-tone information. In contrast with previous ROI detection schemes, our method is easier to be implemented with the existing H.264/AVC

encoder because the ROI detection scheme is easy to derive. When the ROI mask is effectively determined, more resources including bits and computing power are allocated to the ROI. Therefore, several coding parameters (e.g. QP, MB candidate modes, the number of reference frames, motion estimation accuracy and motion estimation search range) should be adaptively adjusted at the MB level according to the ROI mask. In this way, the visual quality of the area which is more sensitive to the HVS is emphasized, hence the overall subjective visual quality is improved. Meanwhile, the decoding complexity is also optimized with an ROI based R-D-C cost function at the encoder side to build a decoding-friendly scheme.

Compared with the latest JVT H.264/AVC reference model JM9.8 [62], our encoder can be used to achieve much better subjective visual quality with comparable PSNR value, and the ROI based rate control can be applied to reduce the mismatch between actual frame bits and target frame bits by up to 80%. Meanwhile, the ME/MC time is saved by up to 73%, the overall coding time is saved by up to 63% and the decoding interpolation complexity is also reduced by up to 48%. Therefore, our simplified H.264/AVC encoder is more effective for real-time conversational video communication.

Chapter 5

Bit Allocation of Rate Control for H.264/AVC Scalable Extension

5.1 Introduction

Highly scalable coding is intended to accommodate decoding for various video applications with diverse needs in resolution, video quality and frame rate. The scalable functions addressed in this chapter include SNR (quality), spatial, temporal and combined scalability. The objective of scalable coding is to achieve high scalability without a significant loss in coding efficiency. Recently, a scalable extension of H.264/AVC proposed by HHI [46] has been adopted as the reference software for scalable video model, which is named joint scalable video model (JSVM). A full spatial-temporal-quality scalable is provided by introducing new concepts of inter-layer prediction and hierarchical B frame, as described in Section 2.3. Inter-layer prediction is employed to remove the redundancy between layers, and the motion and texture information in an enhancement layer is adaptively predicted from that in its base layer. The effective inter-layer prediction of JSVM enables the encoder to achieve scalable property without much coding efficiency lose.

Besides scalability, rate control has also played an important role in video services, especially for the real-time communications. As discussed in Chapter 3, with a proper rate control strategy, the encoding parameters could be adjusted to prevent the buffer from overflowing (causes frame skipping) and underflowing (causes wasting of channel resource). In the current JSVM reference software [46], there is no rate control mechanism for each scalable layer, the fixed bitrate encoding for each scalable layer is realized by two methods [33]. The first one is the FGS approach. In this method, for each scalable layer a minimal bitrate is first generated by a initial QP value, and some FGS enhancement data are next progressively encoded with decreasing quantization parameter:

$$QP[k] = \begin{cases} QP_{initial} & : k = 0 \\ \min(QP[k-1] - 6, 0) & : k > 0 \end{cases} \quad (5.1)$$

where $QP[k]$ is the quantization parameter to generate the k -th enhancement FGS data (the FGS base data is denoted by $k = 0$), $QP_{initial}$ is the initial quantization parameter for the current scalable layer. Higher maximal bitrates could be achieved with more FGS enhancement data, and any target bitrate for each scalable layer between the minimal and maximal bitrates could be extracted by truncating the progressive refinement FGS data. In this way, to generate fixed bitrate bitstreams for n scalable layers, encoder is required to run $(n + 1)$ times (the FGS truncation of previous scalable layer and FGS generation for the current scalable layer could be run simultaneously). Although the FGS refinement and truncation method could achieve the target bitrate, this kind of fixed bitrate encoding does not take the buffer status into consideration, and the FGS approach might significantly decrease the coding efficiency if the bits generated by the predefined quantization step is much smaller than the target bits, because the FGS data

only refines the texture information and the corresponding motion refinement is not included. Thus, we need to study a "Cross SNR layer ME/MC scheme", which improves the coding efficiency by refining the motion information in the enhancement layers. Another approach to realize fixed bitrate coding for each scalable layer is to repeatedly use the logarithmic search algorithm to find a fixed quantization parameter, which might be employed to encode the scalable layer with target bitrate constraint. However, this algorithm tries different quantization parameters multiple times for the whole sequence. The generated bitrate value with each trial quantization parameter is further used as a feedback information to adjust the quantization parameter value for the next try, until the bitrate generated is within the acceptable range or the encoding time exceeds the pre-defined maximal threshold. Apparently, this kind of algorithm that encodes the whole sequence multiple times is not applicable for real-time applications, and the buffer status is not considered either. Therefore, an applicable rate control scheme with buffer control is desired for the scalable video encoder.

In this chapter, our rate control scheme for H.264/AVC presented in Chapter 3 is extended to support its scalable extension. First, we study a cross SNR layer ME/MC scheme which provides a good trade-off between motion and texture information for all SNR layers. Then, the inter-layer property and hierarchical temporal prediction structure are analyzed within our rate control strategy. With our previously proposed adaptive MAD prediction and sum bits R-Q model, both the frame level and MB level rate control could be achieved for H.264/AVC scalable extension.

The rest of this chapter is organized as follows. A cross SNR layer ME/MC scheme is first studied in Section 5.2. It refines the motion information adaptively in the enhancement layer. The rate control scheme for SNR and spatial enhancement

layers is introduced in Section 5.3, where MAD is adaptively predicted from either the temporal information of the current layer or the spatial information of the previous layer, and the actual bits used in the previous layer are also taken into consideration for the target frame bits allocation of the current layer. In Section 5.4, the temporal prediction structure of hierarchical B frame is analyzed for bit allocation, and a corresponding rate control scheme is also proposed. Combining the methods from Section 5.3 and Section 5.4, the rate control scheme for combined enhancement layer is described in Section 5.5. Section 5.6 presents the experimental results and discussions. Finally, the conclusions are given in Section 5.7.

5.2 Cross SNR Layer ME/MC Scheme

To achieve SNR scalability, the trade-off between motion information and residual information should be considered at all SNR layers. By using the rate distortion optimization (RDO) that is formulated as the minimization of the following cost function:

$$J = D + \lambda_{opt}R \quad (5.2)$$

where the distortion D , representing the residual (texture or prediction error), is weighted against the number of bits R associated with the motion information using a Lagrange multiplier λ_{opt} . Each λ_{opt} corresponds to a bitrate range and a tradeoff between motion information and residual information. A large λ_{opt} performs better at a low bitrates while a small λ_{opt} works well at a high bitrates [45]. A large λ_{opt} is normally selected in the existing SNR scalability schemes [45, 105–109] to generate the motion vector field (MVF). Only the residual information is refined at high bitrates [106–109]. Clearly, in these SNR scalability schemes, there is not enough

motion information to support the residual information, and the tradeoff between the motion information and the residual information is very poor at high bitrates. As a result, there still exists a PSNR gap of about 1-3 dB when compared to the state-of-the-art single layer coding scheme at high bitrates [110].

There are some existing works that use the previous base layer information to predict the current enhancement layer frame to increase the FGS coding efficiency [111,112]. The scalable motion information concept combined with wavelet coding was first suggested in [113]. This concept was then used by many scalable coding proposals in the MPEG scalable coding contest in March 2004. In this section, we study the cross layer SNR ME/MC scheme with the concern of RDO and implementation of H.264/AVC scalable extension.

To minimize Eq. (5.2), two parameters are predefined for ME/MC corresponding to different bitrate ranges at the l -th temporal level and the j -th spatial level. Suppose that they are $\lambda_{opt}(l, j)$ and $\lambda_{high}(l, j)$, respectively, and satisfy

$$\lambda_{opt}(l, j) \geq \lambda_{high}(l, j) \quad (5.3)$$

where $\lambda_{opt}(l, j)$ corresponds to the most important bitrate range at the l -th temporal level and the j -th spatial level, which can be determined by the customer composition [114], and $\lambda_{high}(l, j)$ corresponds to high bitrate range.

At the l -th temporal level and the j -th spatial level, ME/MC starts with the parameter $\lambda_{opt}(l, j)$ by using the sum of absolute difference (SAD) $D_p(dx, dy)$:

$$D_p(dx, dy) = \sum_{x,y} |F_1(x, y) - F_2(x - dx, y - dy)| \quad (5.4)$$

where $F_1(x, y)$ is the predicted frame, $F_2(x, y)$ is the reference frame, and (dx, dy) is a motion vector (MV).

5.2 Cross SNR Layer ME/MC Scheme

124

The final MV (dx, dy) is often selected by the encoder to minimize the following performance index:

$$\min\{D_p(dx, dy) + \lambda_{opt}(R_{mv}(dx, dy) + R_{ref}(F_2))\} \quad (5.5)$$

where $R_{mv}(dx, dy)$ and $R_{ref}(F_2)$ are the numbers of bits used to code MV (dx, dy) and the reference index of frame F_2 , respectively [45].

The above ME/MC is called the basic ME/MC. For simplicity, assume that the motion information MV_{opt} and the residual frame $F_1(x, y) - F_2(x - dx_0, y - dy_0)$ ($(dx_0, dy_0) \in MV_{opt}$) are generated by using the basic ME/MC scheme with $\lambda_{opt}(l, j)$.

To maximize the coding efficiency at a high bitrate, both the motion information and the residual information should be refined at the enhancement layers. To achieve this, the coded residual information and motion information MV_{opt} at optimal bitrate base layer should be used by the encoder to generate the motion information and the residual frames at a high bitrate enhancement layer.

Let $Q()$ denote the quantization operation. All motion information MV_{opt} and an optimal part of the residual frame $Q(F_1(x, y) - F_2(x - dx_0, y - dy_0))$ are coded and sent to the decoder at the optimal point. When the bitrate is higher than the optimal point, the remaining energy $D_{br}(dx_0, dy_0, dx_0, dy_0)$ (to be defined in Eq. (5.7)) should be coded and sent to the decoder. There are two possible methods to code $D_{br}(dx_0, dy_0, dx_0, dy_0)$ at a higher bitrate. One way is to code it directly, which is the same as the FGS scheme [105, 115]. Actually, this is the INTRA coding mode of the remaining energy. The other way is to further perform ME/MC to generate a new residual frame and additional MVs, and to code the new residual frame and the additional MVs. This is an INTER coding mode of the remaining energy. Note that if the motion information has been generated under

the assumption that the optimal operating point is at a high bitrate, there will be too much motion information when the optimal bitrate is truncated to a low bitrate, just as in the existing MCTF based SVC schemes [34, 116]. Conversely, MV_{opt} may not be enough at a higher bitrate. Thus, for the enhancement layers at a high bitrate, INTER coding should be more efficient than INTRA coding, especially for the video with slow motion.

The INTER coding is used in the cross SNR layer ME/MC scheme as illustrated in Fig. 5.1. The description of the scheme is given below:

Let (dx_0, dy_0) denote an MV obtained by the basic ME/MC at temporal level l and spatial level j . Define a set of MVs around (dx_0, dy_0) as

$$\rho(dx_0, dy_0, \delta_x, \delta_y) = \{(dx, dy) | |dx - dx_0| < \delta_x, |dy - dy_0| < \delta_y\} \quad (5.6)$$

Suppose that (dx, dy) is in the set $\rho(dx_0, dy_0, \delta_{xrb}, \delta_{yrb})$ and is a candidate for the refinement. Define an SNR refinement criterion as

$$D_{br}(dx, dy, dx_0, dy_0) = \sum_{x,y} |F_1(x, y) - F_2(x - dx, y - dy) - IQ(Q(F_1(x, y) - F_2(x - dx_0, y - dy_0)))| \quad (5.7)$$

where $IQ()$ denotes the inverse quantization operation.

Let $R_{mv}(dx - dx_0, dy - dy_0)$ denote the number of bits to code MV $(dx - dx_0, dy - dy_0)$. A further ME/MC with an MV of (dx, dy) will be performed at the l -th temporal level and the j -th spatial level if the inequality

$$\begin{aligned} & D_{br}(dx, dy, dx_0, dy_0) + \lambda_{high}(l, j) R_{mv}(dx - dx_0, dy - dy_0) \\ & \leq D_{br}(dx_0, dy_0, dx_0, dy_0) \end{aligned} \quad (5.8)$$

is satisfied.

5.2 Cross SNR Layer ME/MC Scheme

126

Figure 5.1: Cross SNR layer ME/MC scheme.

5.2 Cross SNR Layer ME/MC Scheme

127

If there is no (dx, dy) in the set $\rho(dx_0, dy_0, \delta_{xrb}, \delta_{yrb})$ such that Eq. (5.8) is satisfied, then the remaining energy $D_{br}(dx_0, dy_0, dx_0, dy_0)$ is coded directly.

Let (dx_1, dy_1) and (dx_2, dy_2) be two refined MV candidates of (dx_0, dy_0) . (dx_1, dy_1) will be chosen if

$$\begin{aligned} & D_{br}(dx_1, dy_1, dx_0, dy_0) + \lambda_{high}(l, j)R_{mv}(dx_1 - dx_0, dy_1 - dy_0) \\ \leq & D_{br}(dx_2, dy_2, dx_0, dy_0) + \lambda_{high}(l, j)R_{mv}(dx_2 - dx_0, dy_2 - dy_0) \end{aligned} \quad (5.9)$$

Note that the inequality in the existing ME/MC scheme is given by

$$D_p(dx_1, dy_1) + \lambda_{high}(l, j)R_{mv}(dx_1, dy_1) \leq D_p(dx_2, dy_2) + \lambda_{high}(l, j)R_{mv}(dx_2, dy_2) \quad (5.10)$$

where $D_{br}(dx, dy, dx_0, dy_0)$ in Eq. (5.7) is the SAD between the residual information to be encoded in the enhancement layer and the residual information reconstructed from the base layer, and $D_p(dx, dy)$ in Eq. (5.4) is the SAD between the reference frame and the predicted frame. The only difference between $D_{br}(dx, dy, dx_0, dy_0)$ and $D_p(dx, dy)$ is that an additional item $IQ(Q(F_1(x, y) - F_2(x - dx_0, y - dy_0)))$ is involved in $D_{br}(dx, dy, dx_0, dy_0)$. Eq. (5.9) can be regarded as a generalized version of Eq. (5.10).

Similar to the SVC scheme in [45], there are many coding modes for the remaining energy. The mode decision process for the the cross SNR layer ME/MC scheme is the same as that in [45] except that the performance indices of Eqs. (5.4) and (5.10) are replaced by those defined in Eqs. (5.7) and (5.9).

For the sake of simplicity, we let $\Theta(MV, \lambda_{opt}(l, j), l, k, j)$ and $\Theta(MV, \lambda_{high}(l, j), l, k, j)$ denote respectively the motion information to be coded at an optimal bitrate and a high bitrate, while $\Theta(O, RI, l, k, j)$ and $\Theta(H, RI, l, k, j)$ denote respectively the residual frames obtained by using $\lambda_{opt}(l, j)$ and $\lambda_{high}(l, j)$ for the k -th motion compensation pair at temporal level l and spatial level j . We also let $R_{mv}(\Theta(MV, \lambda_{high}(l, j), l, k, j) - \Theta(MV, \lambda_{opt}(l, j), l, k, j))$ denote the number of bits to code all motion information for the k -th

5.2 Cross SNR Layer ME/MC Scheme

128

motion compensation pair at high bitrates, and $D_{br}(\Theta(MV, \lambda_{high}(l, j), l, k, j), \Theta(MV, \lambda_{opt}(l, j), l, k, j))$ denote the SAD of the k -th residual frame that can be computed similar to Eq. (5.7). We need to determine the bitrate at which we would switch the residual frame from $\Theta(O, RI, l, k, j)$ to $\Theta(H, RI, l, k, j)$ when the cross SNR layer ME/MC scheme is used.

The switching bitrate is computed to be

$$\min_{\lambda} \{R(\lambda)\} \quad (5.11)$$

such that

$$\begin{aligned} & D_{br}(\Theta(MV, \lambda_{opt}(l, j), l, k, j), \Theta(MV, \lambda_{opt}(l, j), l, k, j)) \\ & \geq \lambda R_{mv}(\Theta(MV, \lambda_{high}(l, j), l, k, j) - \Theta(MV, \lambda_{opt}(l, j), l, k, j)) + \\ & D_{br}(\Theta(MV, \lambda_{high}(l, j), l, k, j), \Theta(MV, \lambda_{opt}(l, j), l, k, j)) \end{aligned} \quad (5.12)$$

It should be highlighted that both the motion information and the residual frame to be coded are switched when the cross SNR layer ME/MC scheme is used. There may be multiple MVFs in the cross SNR layer ME/MC scheme. Thus, the cross SNR layer ME/MC scheme outperforms the existing SNR scalability schemes without motion refinement [106–109, 111] in the sense that the cross SNR layer ME/MC scheme can be used to achieve a good trade-off between residual information and motion information at all bitrates. Thus, the cross SNR layer ME/MC scheme is utilized to do inter-layer prediction for scalable video coding.

5.3 Bit Allocation of Rate Control for SNR and Spatial Enhancement Layer

In the case of SNR and spatial scalable video coding, all the salable layers are encoded with the same frame rate. The inter-layer prediction, as described in Section 2.3, is used to remove the correlation between the scalable layers. If the enhancement layer is with higher spatial resolution than its base layer, it is named as spatial enhancement layer. If both the enhancement layer and its base layer share the same spatial resolution, the enhancement layer is a coarse gain SNR (CGS) enhancement layer. It should be mentioned that the cross SNR layer ME/MC scheme described in Section 5.2 is proposed in our earlier work independently [117] and similar to the CGS scheme adopted in HHI work [46]. Both our scheme and HHI scheme present the same idea to optimally allocate bits between motion and residual information in the enhancement layer.

The same as our bit allocation scheme presented in Chapter 3, mean absolute difference (MAD) is chosen to indicate the coding complexity in this chapter. In the first layer, the MAD is calculated between the predicted frame and the reference frame. In the enhancement layer, an adaptive inter-layer residual prediction approach is employed. The residual information already coded in the previous base layer can usually be used to predict the residual information to be coded in the enhancement layer, when the motion information in the enhancement layer is similar to the base layer. However, if the motion information in the enhancement layer differs a lot from that in its base layer, the residual information in the base layer may be useless. Therefore, when the MAD is calculated in the enhancement layer, the coded residual information in the base layer should be further subtracted, adaptively, from the difference between the predicted frame and its referencing one.

As discussed in Section 3.2, there is a QP-MAD dilemma caused by the RDO in H.264/AVC. To solve this problem, a simple temporal model is proposed in [21] to predict the MAD before motion compensation, but the predicted MAD by temporal model is

5.3 Bit Allocation of Rate Control for SNR and Spatial Enhancement Layer

130

not accurate when MAD changes abruptly due to high motion or scene changes.

We have proposed an improved MAD prediction scheme in Section 3.3 for the H.264/AVC rate control, and similar idea can be extended to the SNR and spatial enhancement layers. In the enhancement layer, two MAD prediction models are employed at the frame level:

$$MAD_{pred,1}[j][i] = Y1[j] \times MAD_{actual}[j][i-1] + Y2[j] \quad (5.13)$$

$$MAD_{pred,2}[j][i] = Z1[j] \times MAD_{actual}[j-1][i] + Z2[j] \quad (5.14)$$

where i is the frame index and j is the layer index, and $Y1, Y2, Z1, Z2$ are the linear model parameters, which should be updated after encoding each frame. It can be shown that Eq. (5.13) is predicted from the information of the previous frame in the current enhancement layer, and Eq. (5.14) is predicted from the information of the current frame in the previous base layer. Clearly, Eq. (5.14) can be used to achieve more accurate prediction when an abrupt MAD change happens, as the change has already been reflected by the MAD of the current frame in the previous base layer.

As the spatial MAD prediction presented in Eq. (3.13) of Section 3.3, Eq. (5.14) could efficiently predict the abrupt change of MAD, but for some sequences with fine texture details or irregular regional motion, this direct prediction model sometimes does not work as well at the frame level. Here, we could use the same switching policy, as proposed in Section 3.3, to choose the MAD prediction model with higher accuracy. Meanwhile, the MB level MAD prediction could also be achieved in the enhancement layer, $MAD_{pred,1}$ could be calculated based on the co-site MB of the previous frame in the current enhancement layer, and $MAD_{pred,2}$ could be predicted from the co-site MB of the current frame in the previous base layer.

With the above method, our proposed MAD prediction model for SNR and spatial enhancement layers utilizes the available information from both the previous frame of the

current layer and the current frame of the previous base layer. Thus, we can accurately estimate the coding complexity MAD before ME/MC, especially when high motion or scene change happens. The MB level adaptive MAD prediction further improves the prediction efficiency when regional high motion or fine texture occurs.

Meanwhile, the linear sum bits R-Q model, which is described in Section 3.4 to solve the problem caused by inaccurate estimation of texture bits for H.264/AVC, could also be used for the rate control in scalable extension of H.264/AVC. Compared with a quadratic model, the linear R-Q model is much easier for QP calculation and R-Q model training. The the linear sum bits R-Q model is the same as Eq. (3.17).

The same as rate control scheme for H.264/AVC, the QP calculation could be optimized at MB level, as described in Section 3.5.

The overall rate control algorithm for the SNR and spatial enhancement layers could follow the description of Section 3.6, except for the target bit allocation at the frame level. After the target bits are bounded by HRD, as Eq. (3.38), the bits used to code all the previous layers of the current frame should be subtracted from the total number of target frame bits, i.e.

$$R_{sum}[i] = R_{sum}[i] - R_{pre-layer}[i] \quad (5.15)$$

where $R_{pre-layer}[i]$ is the total actual bits used to encode all the previous layers of the current frame.

5.4 Bit Allocation of Rate Control for Hierarchical B Frame

Compared with the previous video coding standards, H.264/AVC supports much more flexibility at the frame level, the coding and display order of frames is completely decoupled in H.264/AVC, and any frame could be used as the reference frame for motion

estimation and compensation of the subsequent frames, independent of its frame type. During the development of techniques for scalable video coding, the coding structure with hierarchical B frame improves the coding efficiency significantly as compared to the traditional “IBBP” structure for a wide range of test sequences [51].

In contrast to the traditional “IPPP” or “IBBP” coding structure, the rate control for hierarchical B frame structure is more challenging, because the bit allocation should take the hierarchical level temporal prediction into consideration. An example of video coding with hierarchical B frame structure is given in Section 2.3.2. In Fig. 2.7, the I_0/P_0 frames are at the lowest hierarchical level, and B_3 frames are at the highest hierarchical level. The frames at the lower hierarchical level are the reference frames for the frames at the next higher hierarchical level (e.g. B_1 is the reference frame for B_2 , and B_2 is the reference frame for B_3). Therefore, more bits should be allocated to the frames at lower hierarchical level to obtain a high quality for these frames, and this could improve the motion prediction efficiency for the following frames at higher hierarchical level. For example, in Fig. 2.7, the I_0/P_0 frames should be encoded with more bits for high quality, because these frames are directly or indirectly used as the reference frames for motion prediction for all the other B frames. At the next higher hierarchical level, B_1 frames should be allocated with relatively fewer bits compared to I_0/P_0 , because they are used as the reference for fewer frames.

To allocate target frame bits according to the hierarchical structure, Eq. (3.33) in Section 3.6 should be changed to:

$$R_2[i] = R_r[i] \times \frac{w_k}{w_{sum}} \quad (5.16)$$

where k is the index of hierarchical level ($k = 0$ denotes the level for I_0/P_0 frames) of the current frame, w_k is the weighting factor for the frames at k -th hierarchical level

and w_{sum} is the sum of weighting factors of the remaining frames, and

$$w_{sum} = \sum_{m=0}^{L_{max}} N_{r,m} \times w_m \quad (5.17)$$

and where L_{max} is the maximal index of hierarchical level, $N_{r,m}$ is the remaining number of the frames at the $m - th$ hierarchical level. The relationship between the weighting factors should satisfy

$$w_{m1} \geq w_{m2} \quad \text{if} \quad m1 < m2 \quad (5.18)$$

For the example in Fig. 5.2, the weighting factors for the 4 hierarchical levels could be set as

$$w_m = \begin{cases} 1.0 & : m = 1 \\ 0.5 & : m = 2 \\ 0.4 & : m = 3 \\ 0.3 & : m = 4 \end{cases} \quad (5.19)$$

Note that the above values are chosen heuristically, they may not be optimal for all the test sequences. Some results in Section 5.6 shows this imperfection. To maximize coding efficiency, it is desirable to set the weighting factors adaptively to different video content and this will be studied in the future work.

Except for the frame level bit allocation scheme, the other steps of rate control for hierarchical B frame are the same as the algorithm described in Section 3.6.

5.5 Bit Allocation of Rate Control for Combined Enhancement Layer

The rate control for the combined enhancement layer of H.264/AVC scalable extension could be regarded as a combination of the rate control for SNR and spatial enhancement layers and the rate control for hierarchical B frame.

Figure 5.2: An example of combined enhancement layer.

An example of combined scalable video coding is shown in Fig. 5.2 [46], where the spatial resolution of the enhancement layer is the up-sampled of the base layer and the temporal resolution is refined with the doubling of the frame rate. Both the inter-layer prediction and the hierarchical structure are employed to encode the combined enhancement layer. Therefore, both the MAD prediction scheme and target frame bit subtraction (Eq. (5.15)) for SNR and spatial enhancement layers described in Section 5.3 and the hierarchical bit allocation scheme described in Section 5.4 should be used for the combined enhancement layer. Note that, for the frames belonging to the refined

temporal resolution, there is no corresponding frames in the base layer as reference to perform the inter-layer prediction, so the target frame bit subtraction is not utilized for these frames.

5.6 Experimental Results

In this section, we shall evaluate the performance of our proposed rate control scheme for H.264/AVC scalable extension. The simulation is implemented with the reference software JSVM7.9 [33]. JSVM7.9 does not provide an efficient rate control scheme, the fixed bitrate coding could be achieved either by FGS layer generation and truncation or by repeatedly using the logarithmic search algorithm to repeatedly encode the whole sequence multiple times. Both these two methods are compared with our rate control scheme in this section. Some standard MPEG test sequences are carried out in our experiments. The rate control scheme for SNR enhancement layer, spatial enhancement layer, hierarchical B frame and combined enhancement layer are tested respectively.

Table 5.1: The Setting of SNR and Spatial Scalable Coding.

Layer No.	Spatial Resolution	Temporal Resolution	Initial QP
0	QCIF	15Hz	40
1	QCIF	15Hz	35
2	CIF	15Hz	32

For SNR and spatial enhancement layer rate control, three scalable layers are simulated as the setting shown in Table 5.1. Layer 0 is the first layer which is encoded without any inter-layer prediction. Layer 1 is a cross gain SNR enhancement layer and Layer 2 is a spatial enhancement layer, both layers are encoded using the adaptive inter-layer prediction from their corresponding base layers (Layer 0 is the base layer of Layer 1, and Layer 1 is the base layer of Layer 2). To compare our scheme with JSVM7.9, the initial setting of QP is the same for both schemes. For real-time application, GOP size is set to 1, and all the test sequences are intra-coded for the first frame (I frame) and followed

5.6 Experimental Results

136

with subsequent inter-coded frames (P frames).

Table 5.2: The Setting of Hierarchical B Frame.

Layer No.	Spatial Resolution	Temporal Resolution	Initial QP
0	QCIF	15Hz	40

For hierarchical B frame rate control, only one layer with hierarchical prediction structure is tested. The testing condition is given in Table 5.2. The GOP size is 4, it means there are two hierarchical prediction level. The key frame (as P_0 in Fig. 2.7) of the current GOP is inter predicted from the key frame of the previous GOP as P frame, only the first frame of the sequence is intra coded.

Table 5.3: The Setting of Combined Scalable Coding.

Layer No.	Temporal Resolution	Spatial Resolution	Initial QP
0	7.5Hz	QCIF	40
1	15Hz	QCIF	40

For combined enhancement layer rate control, two scalable layers are simulated with the setting as shown in Table 5.3. Layer 0 is the first layer which is encoded at 7.5Hz temporal resolution without any inter-layer prediction. In Layer 0, the GOP size is set to 1, all the test sequences are intra-coded for the first frame (I frame) and followed with subsequent inter-coded frames (P frames). Layer 1 is a combined enhancement layer, in which the GOP size is set to 2 and the temporal resolution is 15Hz. In Layer 1, the key frames of the current GOP are inter predicted from the key frames of the previous GOP as P frame, only the first frame of the sequence is intra coded. The bi-directional predicted B frames in Layer 1 are the temporal refinement frames of layer 0, they are encoded without inter-layer prediction. The key frames of Layer 1 are encoded using the adaptive inter-layer prediction from their corresponding frames of layer 0.

We first compare our rate control scheme with FGS method. As shown in Tables 5.4, 5.5 and 5.6, both schemes could generate the target bitrate only with very tiny mismatch. JSVM7.9 uses multiple encoder runs to generate and truncate the FGS data. The target

5.6 Experimental Results

137

Table 5.4: Compare Our Rate Control Scheme with FGS Scheme for SNR and Spatial Scalable Coding.

Sequence	layer No.	tgt. bitr. (kb/s)	rate ctrl.	act. bitr. (kb/s)	PSNR (dB)	gain (dB)
BUS	0	128	no	128.00	27.6	-
			yes	128.07	30.7	+3.1
	1	256	no	255.99	30.4	-
			yes	256.07	33.2	+2.8
	2	1440	no	1439.99	35.3	-
			yes	1441.86	37.8	+2.5
CITY	0	48	no	47.99	27.9	-
			yes	48.02	32.6	+4.7
	1	96	no	95.99	31.1	-
			yes	96.05	34.2	+3.1
	2	480	no	479.99	33.6	-
			yes	480.06	36.7	+3.1
CREW	0	96	no	95.99	31.0	-
			yes	96.02	32.8	+1.8
	1	192	no	191.99	33.4	-
			yes	192.05	35.1	+1.7
	2	1080	no	1079.99	38.8	-
			yes	1080.05	40.0	+1.2
FOOTBALL	0	192	no	191.99	30.6	-
			yes	191.92	32.2	+1.6
	1	384	no	383.99	33.5	-
			yes	383.94	35.7	+2.2
	2	2160	no	2159.99	39.6	-
			yes	2160.06	41.7	+2.1
FOREMAN	0	64	no	63.99	29.5	-
			yes	64.06	33.6	+4.1
	1	128	no	127.99	32.4	-
			yes	128.04	35.6	+3.2
	2	512	no	511.99	35.3	-
			yes	512.08	37.4	+2.1

5.6 Experimental Results

138

Table 5.5: Compare Our Rate Control Scheme with FGS Scheme for Hierarchical B Frame Coding.

Sequence	tgt. bitr. (kb/s)	rate ctrl.	act. bitr. (kb/s)	PSNR (dB)	gain (dB)
BUS	128	no	128.00	30.5	-
		yes	128.11	35.4	+4.9
CITY	48	no	48.00	31.1	-
		yes	48.02	32.6	+1.5
CREW	96	no	95.99	32.9	-
		yes	96.07	33.3	+0.4
FOOTBALL	192	no	192.02	31.7	-
		yes	192.17	37.0	+5.3
FOREMAN	64	no	64.00	32.7	-
		yes	64.20	34.3	+1.6

Table 5.6: Compare Our Rate Control Scheme with FGS Scheme for Combined Scalable Coding.

Sequence	layer No.	tgt. bitr. (kb/s)	rate ctrl.	act. bitr. (kb/s)	PSNR (dB)	gain (dB)
BUS	0	64	no	63.94	28.2	-
			yes	64.04	37.6	+9.4
	1	128	no	127.94	28.7	-
			yes	128.21	35.4	+6.7
CITY	0	24	no	23.98	28.9	-
			yes	24.00	32.2	+3.3
	1	48	no	47.92	29.3	-
			yes	48.05	32.2	+2.9
CREW	0	48	no	47.99	30.9	-
			yes	48.10	33.0	+2.1
	1	96	no	95.96	31.5	-
			yes	96.10	33.2	+1.7
FOOTBALL	0	96	no	95.99	31.8	-
			yes	96.16	40.1	+8.3
	1	192	no	191.96	31.6	-
			yes	192.14	36.9	+5.3
FOREMAN	0	32	no	31.99	30.4	-
			yes	31.99	33.0	+2.6
	1	64	no	63.96	30.9	-
			yes	64.10	33.3	+2.4

5.6 Experimental Results

139

Table 5.7: Compare Our Rate Control Scheme with Fixed QP Scheme for CGS and Spatial Scalable Coding.

Sequence	layer No.	tgt. bitr. (kb/s)	rate ctrl.	act. bitr. (kb/s)	PSNR (dB)	iter.
BUS	0	128	no	132.0	31.1	15
			yes	128.0	30.8	1
	1	256	no	261.9	33.4	15
			yes	256.0	33.1	1
	2	1440	no	1453.9	38.2	6
			yes	1439.7	37.9	1
CITY	0	48	no	48.6	33.1	15
			yes	48.0	32.9	1
	1	96	no	97.6	34.5	15
			yes	96.0	34.2	1
	2	480	no	483.8	36.9	3
			yes	480.0	36.8	1
CREW	0	96	no	104.3	33.4	15
			yes	96.1	33.1	1
	1	192	no	198.5	34.8	15
			yes	192.0	34.6	1
	2	1080	no	1088.7	40.2	4
			yes	1080.0	40.0	1
FOOTBALL	0	192	no	193.9	32.1	11
			yes	191.9	32.1	1
	1	384	no	387.5	34.1	4
			yes	383.9	34.2	1
	2	2160	no	2068.9	41.6	15
			yes	2160.0	41.6	1
FOREMAN	0	64	no	61.4	33.8	15
			yes	64.1	33.7	1
	1	128	no	125.3	35.9	15
			yes	128.0	35.4	1
	2	512	no	507.2	37.7	8
			yes	512.1	37.5	1

5.6 Experimental Results

140

Table 5.8: Compare Our Rate Control Scheme with Fixed QP Scheme for Hierarchical B Frames Coding.

Sequence	tgt. bitr. (kb/s)	rate ctrl.	act. bitr. (kb/s)	PSNR (dB)	iter.
BUS	128	no	128.0	31.9	7
		yes	128.1	31.1	1
CITY	48	no	45.1	33.9	15
		yes	48.0	32.8	1
CREW	96	no	94.1	33.5	15
		yes	96.0	32.3	1
FOOTBALL	192	no	191.0	32.3	5
		yes	192.0	31.7	1
FOREMAN	64	no	64.6	35.0	7
		yes	64.0	33.9	1

Table 5.9: Compare Our Rate Control Scheme with Fixed QP Scheme for Combined Scalable Coding.

Sequence	layer No.	tgt. bitr. (kb/s)	rate ctrl.	act. bitr. (kb/s)	PSNR (dB)	iter.
BUS	0	64	no	58.9	29.0	15
			yes	64.3	29.3	1
	1	128	no	136.1	30.5	15
			yes	128.3	29.5	1
CITY	0	24	no	22.4	31.2	15
			yes	24.1	31.0	1
	1	48	no	48.3	32.3	4
			yes	48.0	31.1	1
CREW	0	48	no	45.0	31.6	15
			yes	48.0	31.7	1
	1	96	no	98.3	32.1	15
			yes	96.0	31.7	1
FOOTBALL	0	96	no	91.4	30.9	15
			yes	96.0	31.2	1
	1	192	no	196.3	30.3	15
			yes	192.1	30.1	1
FOREMAN	0	32	no	33.4	32.3	15
			yes	32.1	31.8	1
	1	64	no	63.8	32.9	5
			yes	64.0	31.7	1

bitrate cannot be achieved if it is less than the bits generated with an initial QP. On the other hand, if the bits generated with an initial QP are much less than the target bits, the coding efficiency will drop, because there is no motion information to support the spatial FGS data. However, to exactly set the QP value for the FGS based fixed bitrate video coding is a very difficult task and may be impossible to achieve when there is no knowledge about the video content to be encoded. On the other hand, our proposed rate control scheme could adjust the QP value at MB level, motion and texture information are balanced well during each MB encoding process. The comparison of coding efficiency is also shown in Tables 5.4, 5.5 and 5.6, where the improvement of luminance PSNR is up to 9.4dB.

We next compare our rate control scheme with fixed QP method. As shown in Tables 5.7, 5.8 and 5.9, our scheme could generate the target bitrate with only very tiny mismatch. On the other hand, the fixed QP encoding with logarithmic QP search sometimes terminates when the iterations reach a predefined threshold, even though the mismatch is larger than expected. In our test with the above setting, the mismatch could be up to 9%. A better guess of the initial QP value could reduce the iteration time, but it depends on the test sequence and other coding parameters. It is difficult to optimally set the initial QP only according to the target bitrate, especially when the content of the sequence is unknown. The comparison of coding efficiency is also shown in Tables 5.7, 5.8 and 5.9. With the fixed QP determined by logarithmic search, JSVM7.9 achieves almost the upper limit of achievable coding efficiency with the same coding parameters. For the spatial and CGS scalable coding, our scheme could achieve comparable coding efficiency. For the hierarchical B frames coding and combined scalable coding, although comparable or even slightly higher PSNR could be achieved for a few layers, the coding efficiency decreases for others. This is because the weighting factors set heuristically in Section 5.4 may not be optimal for all the sequences.

Although the FGS method or fixed QP method provided in JSVM7.9 could achieve target bitrate exactly, the buffer status is not taken into consideration. In our proposed

5.6 Experimental Results

142

Figure 5.3: Buffer status with rate control for SNR and spatial scalable video coding of CREW.

5.6 Experimental Results

143

Figure 5.4: Buffer status with rate control for SNR and spatial scalable video coding of FOOTBALL.

5.6 Experimental Results

144

Figure 5.5: Buffer status with rate control for Hierarchical B frame of BUS and CITY.

5.6 Experimental Results

145

Figure 5.6: Buffer status with rate control for combined scalable video coding of CITY.

5.6 Experimental Results

146

Figure 5.7: Buffer status with rate control for combined scalable video coding of FOREMAN.

scheme, both available bandwidth and buffer status are used to determine target frame bits, so the buffer is well controlled to prevent it from overflowing and underflowing. Meanwhile, the generated frame bits match the target frame bits very well to achieve an accurate bit allocation. The maximal buffer delay is limited to 500ms in the simulation. Due to the accurate bit allocation of our scheme, the buffer is well controlled for all the scalable layers. Some examples of buffer status for SNR enhancement layer, spatial enhancement layer, hierarchical B frame and combined enhancement layer are shown in Figs. 5.3, 5.4, 5.5, 5.6 and 5.7. It is obvious that the encoder with our proposed rate control scheme could efficiently control the buffer status to prevent the buffer from overflowing and underflowing. On the other hand, JSVM7.9 does not provide the buffer control function. Meanwhile, our scheme encodes the sequence only once, while the fixed QP encoding method iterates the coding process for up to 15 times. Our proposed rate control is therefore suitable for real-time applications.

5.7 Conclusions

In this chapter, we have proposed a rate control scheme for the scalable extension of H.264/AVC video coding. Based on the H.264/AVC rate control scheme described in Chapter 3, where the problems associated with the existing H.264/AVC rate control strategy JVT-G012 [21] (inaccurate MAD prediction and inaccurate estimation of the amount of texture bits) are solved by adaptive MAD prediction and sum bits R-Q model, some correlations between scalable layers are explored. This has resulted in the proposal of a bit allocation scheme for scalable extension, which could be utilized for rate control of SNR, spatial, temporal and combined scalable enhancement layers. Compared with the FGS based fixed bitrate video coding, our rate control scheme could also achieve fixed bitrate scalable video coding, and the PSNR is improved by up to 9.4dB due to the good trade-off of motion and residual information in all scalable layers. When compared with the fixed QP encoding method, which iterates the coding process for up to 15 times,

our scheme encodes the sequence only once. So our proposed rate control is suitable for real-time applications. Meanwhile, in our scheme the buffer is well controlled to prevent it from overflowing and underflowing.

Chapter 6

Conclusions and Future Research Directions

6.1 Conclusions

The work within this thesis provides several resource allocation ideas for H.264/AVC and its scalable extension for real-time applications.

Recently, the newly standardized H.264/AVC [42] [22] international video coding standard have attracted many industrial interests. H.264/AVC can perform compression with a bitrate saving of 35-50% when compared to MPEG-4 Visual part 2 and 40-65% when compared to MPEG-2. Furthermore, H.264/AVC has extra advantages such as error resilience and streaming support. All these features make H.264/AVC a very important standard especially in consumer electronics applications. H.264/AVC has been designed for real-time video communication applications [22], such as video conferencing and video telephony as well as to address the needs of video storage (DVD), broadcast video (Cable, DSL, Satellite TV), and video streaming (e.g., video over the Internet, video over wireless) applications. Meanwhile, with the tremendous demands from existing and emerging applications of video transmission over heterogeneous wired/wireless networks, more and more efforts have been invested into researches on scalable video coding (SVC)

schemes in the recent years [12]. SVC schemes are to reliably deliver video to diverse clients over heterogeneous networks using available system resources, particularly for the clients with different display resolutions, caching, intermediate storage resources, access bandwidths, loss rates, or QoS capabilities. The scalable functions provided in the H.264/AVC scalable extension [46] include SNR (quality) scalability, spatial scalability, temporal scalability and combined scalability.

For the real-time video communication applications, it is desirable to produce the best visual quality of the reconstructed video with a given coding resource constraint, so both the bits and the computing power should be properly allocated. In this regards, this thesis provided the following contributions:

- (1) The proposed bit allocation scheme for H.264/AVC rate control outperforms the JVT method by reducing the inaccurate MAD prediction and inaccurate texture bits prediction. It is more suitable for real-time application as the proposed rate control scheme not only enables the encoder to encode the video sequences at a given bitrate, but also controls the buffer much better to prevent its overflowing and underflowing.
- (2) The region-of-interest (ROI) based resource allocation scheme for H.264/AVC standard enables the encoder to allocate more encoding resources (e.g. bandwidth and computing power) to the area with higher visual attention. The ROI information can also be used to build a decoding-friendly encoder with a rate-distortion-complexity (R-D-C) cost function. In this way, the visual quality is improved according to human visual interest and the coding complexity is reduced significantly.
- (3) The proposed bit allocation algorithm for H.264/AVC is extended to H.264/AVC scalable extension. Thus, rate control for all the SNR, spatial, and temporal enhancement layers is achieved.

Compared with the existing video coding schemes, the proposed improvements could achieve better subjective and objective visual quality. Meanwhile, the coding complexity

of video coding system is significantly reduced, and the buffer is well controlled for low delay real-time applications.

6.2 Future Research Directions

Some possible directions of future research are listed as follows:

(a) **ROI based resource allocation for H.264/AVC scalable extension:**

Although we proposed an ROI based resource allocation for H.264/AVC in Chapter 4, the same issue of H.264/AVC scalable extension should also be studied. There might be three possible research problems for this part:

- (1) Due to the QP-motion dilemma of H.264/AVC, as described in Section 4.2.1, the motion information cannot be easily used for fast ROI detection. Detecting ROI without the motion information may result in the loss of some attracting moving area from ROI. However, at the scalable enhancement layers, the motion information of the previous layers are already available. These motion information may help to detect the ROI of the enhancement layer. When taking the motion information into consideration, an ROI detection scheme could be utilized in different kinds of videos (only conversational video is studied in Chapter 4).
- (2) Based on the bit allocation scheme of rate control for H.264/AVC scalable extension, as described in Chapter 5, the ROI information of enhancement layer could help to allocate more bits to the ROI area during scalable refinement. The detailed adjustment of ROI based bit allocation might be different for SNR, spatial and temporal enhancement layers.
- (3) With the detected ROI information of the enhancement layer, the computing power could be accordingly saved for the area outside the ROI. For example, a fast mode decision scheme for scalable enhancement layer might be derived in the future.

(b) Multiple base layers for H.264/AVC scalable extension:

Figure 6.1: An example of multiple base layers.

Let us consider the combined SNR and spatial scalability that is illustrated in Fig. 6.1. There are four layers, (QCIF, Low), (QCIF, Medium), (CIF, Low) and (CIF, Medium). Currently, the H.264/AVC scalable extension only allows each frame at the enhancement layer to have one base layer [33]. Because the quality of base layer should be lower than that of the enhancement layer, (QCIF, Low) should be the “base layer” for both (CIF, Low) and (QCIF, Medium). When the scalable bitstream is generated, the spatial redundancy between (QCIF, Low) and (CIF, Low) and the SNR redundancy between (QCIF, Low) and (QCIF, Medium) can be removed by the inter-layer prediction [33]. However, there is a problem when (CIF, Medium) is coded. If there is only one “base layer” for (CIF, Medium), either (CIF, Low) or (QCIF, Medium) could be chosen as the “base layer”. On one hand, when the (CIF, Low) is chosen as the “base layer”, the SNR redundancy between (CIF, Low) and (CIF, Medium) can be efficiently removed, but the spatial redundancy between (CIF, Medium) and (QCIF, Medium) cannot be removed. On the other hand, when the (QCIF, Medium) is chosen as the “base layer”, the spatial redundancy between

(QCIF, Medium) and (CIF, Medium) can be efficiently removed, but the SNR redundancy between (CIF, Medium) and (CIF, Low) cannot be removed. This implies that the coding efficiency of (CIF, Medium) cannot be optimized if only one “base layer” is used for inter-layer prediction. To solve this problem, each enhancement layer may have more than one “base layer”, (CIF, Medium) could use (QCIF, Medium) and (CIF, Low) as “base layers”. In this way, the motion information and the residual information at the “base layers” could be efficiently utilized when the enhancement layer is coded. The objective is to remove the redundancy among different layers as much as possible. There might be two interesting problems on this point. One is on the design of context models for the entropy coding. The “base layer” information should be used when the context models are designed at the enhancement layers. The other one is on the compensated prediction at the enhancement layer. The base layer information should be fully utilized to remove both the spatial redundancy and the SNR redundancy between the enhancement layer and its “base layers”.

(c) **Truncation scheme for scalable video coding:**

Consider two successive layers: Layer 1 and Layer 2, where Layer 1 is the base layer of Layer 2. There are two possible ways to generate an embedded bitstream including the information of Layers 1 and 2:

- (1) If Layer 1 is more important than Layer 2. The motion information at Layer 1 is first obtained by the RDO of without any constraint, and the motion information of Layer 2 can be refined based on the information of Layer 1. Since the coding process at Layer 2 does not affect the coding efficiency at Layer 1, the coding efficiency at Layer 1 is optimized. This is a refinement method.
- (2) If Layer 2 is more important. The motion information of Layer 2 is first generated by the RDO without any constraint, then the motion information of Layer 1 is truncated from that at Layer 2 with properly defined constraint. In this way, the coding efficiency at Layer 2 is improved as compared to the refinement method.

This is a truncation method.

The refinement method has been well studied by in JSVM [33]. There are some proposed truncation methods in [34,118,119]. In these schemes, motion information and residual information are generated at high bitrates. To obtain a better tradeoff between motion information and residual information at low bitrates, motion information is truncated. However, the residual information coded at low bitrates is still the same one generated at high bitrates. In this way, there is a mismatch between motion information and residual information at low bitrates by existing truncation schemes [34,118,119]. Therefore, to improve the coding efficiency at high bitrates and to minimize the coding efficiency drop at low bitrates, a proper truncation method scheme should be studied in the future.

Author's Publications

Journal Paper:

- Y. Liu, Z. G. Li, and Y. C. Soh, "A novel rate control scheme for low delay video communication of H.264/AVC standard," *IEEE Trans. Circuits, Syst., Video Technol.*, Vol 17, No. 1, Jan, 2007.
- Y. Liu, Z. G. Li, and Y. C. Soh, "Rate control for scalable extension of H.264/AVC," *IEEE Trans. Circuits, Syst., Video Technol.*, Vol 18, No. 1, Jan, 2008.
- Y. Liu, Z. G. Li, and Y. C. Soh, "Region-of-interest based resource allocation for conversational video communication of H.264/AVC," *IEEE Trans. Circuits, Syst., Video Technol.*, Vol 18, No. 1, Jan, 2008.

Conference Paper:

- Y. Liu, Z. G. Li, and Y. C. Soh, "Rate control for spatial/CGS scalable extension of H.264/AVC," *IEEE International Symposium on Circuits and Systems (ISCAS)*, New Orleans, USA, May 2007.
- Y. Liu, Z. G. Li, and Y. C. Soh, "Adaptive MAD prediction and refined R-Q model for H.264/AVC rate control," in *Proc. IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, Toulouse, France, May 2006.

-
- Y. Liu, Z. G. Li, and Y. C. Soh, "Conversational video communication of H.264/AVC with region-of-interest concern," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Atlanta, USA, October 2006.
 - Z. G. Li, Y. Liu, and Y. C. Soh, "A novel SNR refinement scheme for scalable video coding," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Genova, Italy, September 2005, pp. III – 644–7.
 - Y. Liu, Z. G. Li, and Y. C. Soh, "Motion compensated temporal filtering with optimal temporal distance between each motion compensation pair," in *Proc. IEEE International Conference on Image Processing (ICIP)*, Singapore, October 2004, pp. 2283–2286.

Bibliography

- [1] J. Chen, U.-V. Koc, and K. J. R. Liu, *Design of Digital Video Coding Systems*. Marcel Dekker, 2001.
- [2] J. L. Mitchell, W. B. Pennebaker, C. E. Fogg, and D. J. L. Gall, *MPEG Video Compression Standard*. Chapman and Hall, 1996.
- [3] S. J. Choi and J. W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. on Image Processing*, vol. 8, no. 2, pp. 155–167, February 1999.
- [4] J. R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Trans. on Image Processing*, vol. 3, no. 9, pp. 559–571, September 1994.
- [5] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*. New York: Academic, 1990.
- [6] L. Debnath, *Wavelet Transform and Their Applications*. Springer, Hardcover, 2000.
- [7] M. Nelson and J.-L. Gailly, *The Data Compression Book*. Ed. M&T Books, 1996.
- [8] I. Witten, R. Neal, and J. Cleary, "Arithmetic coding for data compression," *Communications of the ACM*, vol. 30, pp. 520–540, 1997.

-
- [9] A. G. and R. M. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic Publisher, 1992.
 - [10] B. G. Haskell, A. Puri, and A. N. Netravali, *Digital Video: An introduction to MPEG-2*. Chapam Hall, 1997.
 - [11] "Video coding for low bitrate communication," ITU-T Recommendation H.263, November 1995.
 - [12] "Applications and requirements for scalable video coding," ISO/IEC JTC1/SC29/WG11/N5880, Trondheim, Norway, July 2003.
 - [13] "Application and requirements for scalable video coding," ISO/IEO JTC1/SC29/WG11/N5880, Trondheim, Norway, July 2003.
 - [14] "MPEG-2 test model 5," Doc. ISO/IEC JTC1/SC29 WG11/93-400, April 1993.
 - [15] J. R-Corbera and S. Lei, "Rate control in DCT video coding for low-delay communication," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 9, no. 1, pp. 172–185, February 1999.
 - [16] "Appendix II of H.263 version 3, test model 12," JVT Tenth Meeting, Portland, OR, USA, August 2000.
 - [17] "H.26L test model long term number 9 (TML-9) draft 0," ITU-T, Video Coding Expert Group (VCEG), December 2001.
 - [18] H. J. Lee, T. H. Chiang, and Y.-Q. Zhang, "Scalable rate control for MPEG-4 video," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 10, pp. 878–894, September 2000.
 - [19] A. Vetro, H. Sun, and Y. Wang, "MPEG-4 rate control for multiple video objects," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 9, pp. 186–199, February 1999.

- [20] "MPEG-4 video verification model v8.0," ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Associated Audio MPEG97/N1796, Stockholm, Sweden, July 1997.
- [21] Z. G. Li, F. Pan, K. P. Lim, G. Feng, X. Lin, and S. Rahardja, "Adaptive basic unit layer rate control for JVT," JVT-G012-rl, 7th Meeting, Pattaya II, Thailand, March 2003.
- [22] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 13, no. 7, July 2003.
- [23] T. Wiegand, F. L. H. Schwarz, A. Joch, and G. J. Sullivan, "Rate constrained coder control and comparison of video coding standards," *IEEE Trans. Circuits, Syst., Video Technol.*, July 2003.
- [24] Y. Liu, Z. G. Li, and Y. C. Soh, "A novel rate control scheme for low delay video communication of H.264/AVC standard," *IEEE Trans. Circuits, Syst., Video Technol.* accepted as full paper, pp 68-78, Vol 17, No. 1, Jan, 2007.
- [25] C. Grecos and M. Y. Yang, "Fast inter mode prediction for p slices in the H.264 video coding standard," *IEEE Trans. Broadcasting*, vol. 51, no. 2, June 2005.
- [26] K. P. Lim, S. Wu, D. J. Wu, S. Rahardja, X. Lin, F. Pan, and Z. G. Li, "Fast INTER mode selection," JVT-I020, San Diego, USA, September 2003.
- [27] Z. Zhou and M.-T. Sun, "Fast macroblock inter mode decision and motion estimation for H.264/MPEG-4 AVC," in *Proc. IEEE International Conference on Image Processing*, September 2004.
- [28] Y.-H. Kim, J.-W. Yoo, S.-W. Lee, J. Shin, J. Paik, and H.-K. Jung, "Adaptive mode decision for H.264 encoder," *ELECTRONICS LETTERS*, vol. 40, no. 19, September 2004.

- [29] Q. Dai, D. Zhu, and R. Ding, "Fast mode decision for inter prediction in H.264," in *Proc. IEEE International Conference on Image Processing*, September 2004.
- [30] P. Yin, H. Tourapis, A. Tourapis, , and J. Boyce, "Fast mode decision and motion estimation for JVT/H.264," in *Proc. IEEE International Conference on Image Processing*, September 2003.
- [31] J. Xin, V. A. and H. Sun, "Efficient macroblock coding-mode decision for H.264/AVC video coding," in *Proc. Picture Coding Symposium*, December 2004.
- [32] X. Lu, A. Tourapis, P. Yin, and J. Boyce, "Fast mode decision and motion estimation for H.264 with a focus on MPEG-2/H.264 transcoding," in *Proc. 2005 IEEE International Symposium on Circuits and Systems*, Kobe, Japan, May 2005.
- [33] "Joint scalable video model JSVM 1.0 software package," Joint Video Team, Doc. JVT-N024, Hong Kong, China, January 2005.
- [34] J. Xu, R. Xiong, B. Feng, G. Sullivan, M. Chieh, F. Wu, and S. Li, "3D sub-band video coding using barbell lifting," ISO/IEC JTC1/SC29 WG11 MPEG2004/M10569/S05, Munchen, Germany, March 2004.
- [35] B. P. Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," in *Proc. IEEE International Conference on Image Processing*, Thessaloniki, Greece, October 2001, pp. 1793–1796.
- [36] A. Secker and D. Tubman, "Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting," in *Proc. IEEE International Conference on Image Processing*, Thessaloniki, Greece, October 2001, pp. 1029–1032.
- [37] "Video codec for audiovisual services at p x 64 kbit/s," CCITT Recommendation H.261, CCITT doc COM XV-R 37-E, 1995.

- [38] "Coding of moving pictures and associated audio for digital storage media at up to 1.5 mbps," ISO/IEC CD 11172, December 1991.
- [39] "Generic coding of moving pictures and associated audio," ISO/IEC CD 13818-22, November 1993.
- [40] S. Battista, F. Casalino, and C. Lande, "A multimedia standard for the third millennium," *IEEE Multimedia*, vol. 6, no. 4, pp. 74–83, 1997.
- [41] N. Brady, "MPEG-4 standardized methods for the compression of arbitrarily shaped video objects," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 9, no. 8, pp. 1170–1189, 1999.
- [42] "Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T REC. H.264 — ISO/IEC 14496-10 AVC)," ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Associated Audio MPEG97/N1796 Doc. JVT-G050r1, Geneva, Switzerland, May 2003.
- [43] R. Schafer, T. Wiegand, and H. Schwarz, "The emerging H.264/AVC standard," *EBU Technical Review*, vol. 1, pp. 253–255, 2003.
- [44] "Overview of H.264/MPEG-4 part10," http://ace.kunsan.ac.kr/seminar/H.264_slides_Rao.pdf.
- [45] H. Schwarz, D. Marpe, and T. Wiegand, "Scalable extension of H.264/AVC," ISO/IEC JTC1/SC29 WG11 MPEG2004/M10569/S03, Munchen, Germany, March 2004.
- [46] "Joint scalable video model JSVM 1.0," Joint Video Team, Doc. JVT-N023, Hong Kong, China, January 2005.
- [47] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 301–317, Mar. 2001.

- [48] M. van der Schaar and H. Radha, "A hybrid temporal-SNR Fine-Grained Scalability for Internet Video," *IEEE Trans. on Circuits and Systems for Video Technology*, 2001.
- [49] H. Radha, M. van der Schaar, and Y. Chen, "The MPEG-4 Fine-Grained Scalable Video Coding method for Multimedia Streaming over IP," *IEEE Tans. on Multimedia*, 2001.
- [50] F. Wu, S. Li, and Y. Q. Zhang, "DCT-prediction based progressive fine granularity scalability," in *Proc. IEEE International Conference on Image Processing*, vol. 3, October 2000, pp. 556–559.
- [51] "Hierarchical B pictures," Joint Video Team, Doc. JVT-P014, Poznan, Poland, July 2005.
- [52] R. Schafer, H. Schwarz, D. Marpe, T. Schierl, and T. Wiegand, "Mctf and scalability extension of h.264/avc and its application to video transmission, storage, and surveillance," in *Proceedings of IS&T/SPIE Symposium on Visual Communications and Image Processing (VCIP'05)*, Beijing, China, July 2005.
- [53] J. Max, "Quantizing for minimum distortion," *IRE Trans. Information Theory*, vol. IT-6, March 1960.
- [54] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*. Englewood Cliffs, NJ: Prentice Hall, 1984.
- [55] A. Ortega, K. Ramchandran, and M. Vetterli, "Optimal trellis based buffered compression and fast approximations," *IEEE Trans. Image Processing*, vol. 3, pp. 26–40, January 1994.
- [56] W. Ding, "Joint encoder and channel rate control of VBR video over ATM networks," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 7, pp. 266–278, April 1997.

- [57] “A MAD based rate control strategy,” Doc. JVT-D070, Klagenfurt, Austria, July 2002.
- [58] C. Wong, O. C. Au, B. Meng, and H. Lam, “Novel H.26X optimal rate. control for low-delay communications,” in *Proc. ICICS-PCM*, Singapore, December 2003, pp. 90–94.
- [59] Z. He, Y. Kim, and S. K. Mitra, “Low-delay rate control and smoothing for video coding via ρ -domain source modeling,” *IEEE Trans. on Circuits and Systems for Video. Technology*, vol. 11, pp. 928–940, August 2001.
- [60] Z. He and S. K. Mitra, “A unified rate-distortion analysis framework for transform coding,” *IEEE Trans. on Circuits and Systems for Video. Technology*, vol. 11, pp. 1221–1236, December 2001.
- [61] —, “A linear source model and a unified rate control algorithm for DCT video coding,” *IEEE Trans. on Circuits and Systems for Video. Technology*, vol. 12, pp. 970–982, November 2002.
- [62] “JM9.8 software package,”
<http://ftp3.itu.ch/av-arch/jvt-site/reference-software/jm98.zip>.
- [63] Z. G. Li, W. Gao, F. Pan, S. W. Ma, K. P. Lim, G. N. Feng, X. Lin, S. Rahardja, H. Q. Lu, and Y. Lu, “Adaptive rate control for H.264,” *Journal of Visual Communication and Image Representation*, vol. 17, no. 2, April 2006.
- [64] M. Jiang and N. Ling, “On lagrange multiplier and quantizer adjustment for H.264 frame-layer video rate control,” *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 16, no. 5, May 2006.
- [65] S. Ma, W. Gao, and Y. Lu, “Rate-distortion analysis for H.264/AVC video coding and its application to rate control,” *IEEE Trans. on Circuits and Systems for Video. Technology*, vol. 15, pp. 1533–1544, December 2005.

- [66] D. A. Pierre, *Optimization Theory with Application*. New York: Dover, 1986.
- [67] H. Yu, Z. Lin, and F. Pan, "An improved rate control algorithm for h.264," in *Proceedings of the 2004 International Symposium on Circuits and Systems*, May 2005.
- [68] J. M. Boyce, "Weighted prediction in the h.264/mpeg avc video coding standard," in *Proceedings of the 2004 International Symposium on Circuits and Systems*, May 2000.
- [69] H. H.-M and C. J.-J, "Source model for transform video coder and its application -part I: Fundamental theory," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 7, April 1997.
- [70] H. S. Malvar, A. Hallapuro, M. Karczewicz, and L. Kerofsky, "Low-complexity transform and quantization in H.264/AVC," *IEEE Trans. Circuits, Syst., Video Technol.*, July 2003.
- [71] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice Hall, 1971.
- [72] J. R. Corbera, P. A. Chou, and S. L. Regunathan, "A generalized hypothetical reference decoder for H.264/AVC," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 13, no. 7, pp. 674–687, July 2003.
- [73] K.-K. Ma and P. I. Hosur, "Performance report of motion vector field adaptive search technique (MVFAST)," in *the 51th MPEG Meeting (ISO/IEC/JTC1/SC29/WG11), M5851*, Noordwijkerhout, Netherlands, March 2000.
- [74] H. S. Bashinski and V. R. Bacharach, "Enhancement of perceptual sensitivity as the result of selectively attending to spatial locations," *Perception Psychopys*, vol. 28, no. 3, pp. 241–280, 1980.

- [75] V. M. Ciaramitaro, E. L. Cameron, and P. Glimcher, "Stimulus probability directs spatial attention: an enhancement of sensitivity in humans and monkeys," *Vision Research*, vol. 41, no. 1, pp. 57–75, January 2001.
- [76] H. Muller and J. M. Findlay, "Sensitivity and criterion effects in the spatial cueing of visual attention," *Perception Psychophys*, vol. 42, no. 4, pp. 383–399, 1987.
- [77] S. Daly, *The Visible Differences Predictor: An Algorithm for The Assessment of Image Fidelity*. Cambridge, MA: MIT Press, 1983.
- [78] S. A. Karunasekera and N. G. Kingsbury, "A distortion measure for blocking artifacts in image based on human visual sensitivity," *IEEE Trans. Image Processing*, June 1995.
- [79] J. Malo, J. Gutierrez, I. Epifanio, F. J. Ferri, and J. M. Artigas, "Perceptual feedback in multigrid motion estimation using an improved DCT quantization," *IEEE Trans. Image Processing*, October 2001.
- [80] S. Winkler, "Vision models and quality metrics for image processing applications," Ph.D. dissertation, Signal Process. Lab., Ecole Polytechnique Federale De Lausanne (EPFL), Swiss Fed. Inst. Technol., December 2000.
- [81] A. B. Watson, J. Hu, and J. F. M. III, "DVQ: A digital video quality metric based on human vision," *J. Elect. Imag.*, January 2001.
- [82] H. Yee, S. Pattanaik, and D. P. Greenberg, "Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments," *ACM Trans. Graphics*, pp. 39–65, 2001.
- [83] S.-S. Kuo and J. D. Johnston, "Spatial noise shaping based on human visual sensitivity and its application to image coding," *IEEE Trans. Image Process.*, May 2002.

-
- [84] N. Doulamis, A. Doulamis, D. Kalogeras, and S. Kollias, "Low bit-rate coding of image sequences using adaptive regions of interest," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 8, no. 8, December 1998.
 - [85] Z. Bojković and D. Milovanović, "Multimedia coding using adaptive regions of interest," in *Proc. 7th Seminar on Neural Networks Applications in Electrical Engineering*, Serbia and Montenegro, September 2004.
 - [86] M.-C. Chi, M.-J. Chen, and C.-T. Hsu, "Region-of interest video coding by fuzzy control for H.263+ standards," in *Proc. IEEE International Symposium on Circuits and Systems*, May 2004.
 - [87] Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao, "Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation," *IEEE Trans. Image Processing*, vol. 14, no. 11, November 2005.
 - [88] W. James, *The Principles of Psychology*. Cambridge, MA: Harvard University Press, 1890.
 - [89] T. Horowitz and A. Treisman, "Attention and apparent motion," *Spatial Vis.*, vol. 8, no. 2, 1994.
 - [90] D. Alais and R. Blake, "Neural strength of visual attention gauged by motion adaptation," *Nature Neurosci.*, vol. 2, no. 11, November 1999.
 - [91] M. J. Black and P. Anandan, "The robust estimation of multiple motions: parametric and piecewise-smooth flow fields," *Comput. Vis. Image Understand.*, vol. 63, no. 1, January 1996.
 - [92] K. Zhang and J. Kittler, "Global motion estimation and robust regression for video coding," in *Proc. IEEE International Conference on Acoustics, Speech, Signal Processing*, 1998.

-
- [93] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 1, January 1998.
- [94] J. Terrillon and S. Akamatsu, "Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images," in *Proc. International Conf on Face and Gesture Recognition*, 2000.
- [95] L. M. Bergasa, M. Mazo, A. Gardel, M. A. Sotelo, and L. Boquete, "Unsupervised and adaptive. gaussian skin-color model," *Image and Vision. Computing*, vol. 18, pp. 987–1003, 2000.
- [96] C. Poynton, *A Technical Introduction to Digital Video*. John Wiley & Sons, January 1996.
- [97] E. Saber and A. M. Tekalp, "Frontal-view face detection and facial feature extraction using color, shape, symmetry based cost functions," *Pattern Recognition Letters*, vol. 19, pp. 669–680, 1998.
- [98] B. Menser and M. Brunig, "Human faces in color images with complex background," *Intelligent Signal Processing and Communications Systems*, pp. 533–536, December 1999.
- [99] K. Sobottka and I. Pitas, "A novel method for automatic face segmentation, facial feature extraction and tracking," *Signal processing: Image communication*, vol. 12, pp. 263–281, 1998.
- [100] R. L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, May 2002.
- [101] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, no. 1, March 2002.

-
- [102] I. E. G. Richardson, *H.264 and MPEG - 4 Video Compression: Video Coding for Next Generation Multimedia Subject Digital Video*. Wiley and Sons, December 2003.
 - [103] V. Lappalainen, A. Hallapuro, and T. D. Hamalainen, "Complexity of optimized H.26L video decoder implementation," *IEEE Trans. Circuits, Syst., Video Technol.*, July 2003.
 - [104] Z. He, Y. Liang, L. Chen, I. Ahmad, and D. Wu, "Power-rate-distortion analysis for wireless video communication under energy constraints," *IEEE Trans. Circuits, Syst., Video Technol.*, May 2005.
 - [105] H. C. Huang, W. H. Peng, Y. C. Lin, C. N. Wang, T. Chiang, and H. M. Hang, "Response of CFP on scalable video coding technology: Proposal S07-A robust scalable video coding," ISO/IEC JTC1/SC29 WG11 MPEG2004/M10569/S07, Munchen, Germany, March 2004.
 - [106] F. Wu, S. P. Li, and Y. Zhang, "A framework for efficient progressive fine granularity scalable video coding," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 11, pp. 332–344, 2001.
 - [107] Y. Zhou, X. Sun, F. Wu, H. Bao, , and S. P. Li, "Flexible p-picture (FLEXP) coding for the efficient fine granular scalability (FGS)," in *Proc. IEEE International Conference on Image Processing*, Singapore, October 2004, pp. 2071–2074.
 - [108] H. Schwarz, D. Marpe, and T. Wiegand, "SNR-scalable extension of H.264/AVC," in *Proc. IEEE International Conference on Image Processing*, Singapore, October 2004, pp. 3113–3116.
 - [109] "Scalable video model 2.0," ISO/IEC JTC 1/SC 29/WG 11/N6520, Redmond, WA, USA, July 2004.

- [110] Y. Wu, A. Golwelkar, and J. W. Woods, "Information technology-coding of audio-visual objects-part 10: Advance video coding," ISO/IEC FDIS 14496-10.
- [111] H. C. Huang, C. N. Wang, and T. Chiang, "A robust fine granularity scalability using trellis-based predictive leak," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 12, pp. 372–384, 2002.
- [112] H.-C. Huang and T. Chiang, "Stack robust fine granularity scalability." in *ISCAS (3)*, 2004, pp. 829–832.
- [113] S. S. Tsai and H.-M. Hang, "Motion information scalability for mc-ezbc," *Signal Processing: Image Communication*, vol. 19, Aug 2004.
- [114] Z. G. Li, Y. C. Soh, and C. Y. Wen, *Switched and Impulsive Systems: Analysis, Design and Application*. Springer-Verlag, 2004.
- [115] X. Sun, Y. Zhou, Y. Wang, G. Sullivan, M. C. Lee, F. Wu, and S. Li, "Progressive fine granularity scalable (PFGS) video coding," ISO/IEC JTC1/SC29 WG11 MPEG2004/M10569/S06, Munchen, Germany, March 2004.
- [116] Y. Wu, A. Golwelkar, and J. W. Woods, "MC-EZBC video proposal from Rensselaer polytechnic institute," ISO/IEC JTC1/SC29 WG11 MPEG2004/M10569/S15, Munchen, Germany, March 2004.
- [117] Z. G. Li, Y. Liu, and Y. C. Soh, "A novel SNR refinement scheme for scalable video coding," in *Proc. IEEE International Conference on Image Processing*, Genova, Italy, September 2005, pp. III – 644–7.
- [118] "Motion information scalability for MC-EZBC: Response to call for evidence on scalable video coding," ISO/IEC JTC1/Sc29/WG11, MPEG2003/m9756, Trondheim, Norway, July 2003.
- [119] "MC-EZBC video proposal from Rensselaer Polytechnic Institute," ISO/IEC JTC1/SC29 WG11 MPEG2004/M10569/S15, Munchen, Germany, March 2004.