Computational Complexity Optimization on H.264 Scalable/Multiview Video Coding

By

Guangyao Zhang

A thesis submitted in partial fulfilment for the requirements for the degree of PhD, at the University of Central Lancashire

April 2014

Student Declaration

Concurrent registration for two or more academic awards

I declare that while registered as a candidate for the research degree, I have not been a registered candidate or enrolled student for another award of the University or other academic or professional institution.

Material submitted for another award

I declare that no material contained in the thesis has been used in any other submission for an academic award and is solely my own work.

Collaboration

This work presented in this thesis was carried out at the ADSIP (Applied Digital Signal and Image Processing) Research Centre, University of Central Lancashire. The work described in the thesis is entirely the candidate's own work.

Signature of Candidate _____

Type of Award Doctor of Philosophy

School School of Computing, Engineering and Physical Sciences

Abstract

The H.264/MPEG-4 Advanced Video Coding (AVC) standard is a high efficiency and flexible video coding standard compared to previous standards. The high efficiency is achieved by utilizing a comprehensive full search motion estimation method. Although the H.264 standard improves the visual quality at low bitrates, it enormously increases the computational complexity. The research described in this thesis focuses on optimization of the computational complexity on H.264 scalable and multiview video coding.

Nowadays, video application areas range from multimedia messaging and mobile to high definition television, and they use different type of transmission systems. The Scalable Video Coding (SVC) extension of the H.264/AVC standard is able to scale the video stream in order to adapt to a variety of devices with different capabilities. Furthermore, a rate control scheme is utilized to improve the visual quality under the constraints of capability and channel bandwidth. However, the computational complexity is increased. A simplified rate control scheme is proposed to reduce the computational complexity. In the proposed scheme, the quantisation parameter can be computed directly instead of using the exhaustive Rate-Quantization model. The linear Mean Absolute Distortion (MAD) prediction model is used to predict the scene change, and the quantisation parameter will be increased directly by a threshold when the scene changes abruptly; otherwise, the comprehensive Rate-Quantisation model will be used. Results show that the optimized rate control scheme is efficient on time saving.

Multiview Video Coding (MVC) is efficient on reducing the huge amount of data in multiple-view video coding. The inter-view reference frames from the adjacent views are exploited for prediction in addition to the temporal prediction. However, due to the increase in the number of reference frames, the computational complexity is also increased. In order to manage the reference frame efficiently, a phase correlation algorithm is utilized to remove the inefficient inter-view reference frame from the reference list. The dependency between the inter-view reference frame and current frame is decided based on the phase correlation coefficients. If the inter-view reference frame is highly related to the current frame, it is still enabled in the reference list; otherwise, it will be disabled. The experimental results show that the proposed scheme is efficient on time saving and without loss in visual quality and increase in bitrate.

The proposed optimization algorithms are efficient in reducing the computational complexity on H.264/AVC extension. The low computational complexity algorithm is useful in the design of future video coding standards, especially on low power handheld devices.

CONTENTS

Abstract I
CONTENTS I
List of Figures V
List of Tables VII
Acknowledgements VIII
List of Abbreviations IX

CHAPTER 1 1
INTRODUCTION 1
1.1 Motivation 1
1.2 History of Video Compression 2
1.3 Concepts and Definitions 5
1.4 Research Objectives 6
1.5 Research Outline and Contributions 7
1.6 Organisation of the Thesis 8

CHAPTER 2 9
AN OVERVIEW OF VIDEO CODING 9
2.1 Introduction 9
2.2 Overview of Video Compression 10
2.3 H.264/MPEG Part 10 Standard 12
2.3.1 Introduction 12
2.3.2 Concept and Definition 12
2.3.3 Inter Prediction 16
2.3.4 Intra Prediction 18
2.3.5 Transform Coding 20
2.3.6 Quantisation Stage 21
2.3.7 Entropy Coding 22
2.4 Summary 23

CHAPTER 3 24
MOTION ESTIMATION STUDY ON H.264/AVC 24
3.1 Introduction 24

3.2 Motion Estimation in H.264/AVC 25
3.2.1 Criterion for Best Match 25
3.2.2 Image Quality 26
3.3 Fast Search Algorithms 27
3.3.1 Four-step Search Method 28
3.3.2 Diamond Search Method 29
3.3.3 Phase Correlation Method 30
3.4 Experimental Results 31
3.5 Summary 35

CHAPTER 4 36
A SIMPLIFIED RATE CONTROL ALGORITHM FOR H.264/SVC 36
4.1 Introduction 36
4.2 H.264/ Scalable Video Coding 37
4.2.1 The Requirements for Scalable Video Coding 37
4.2.2 The Scalability in Scalable Video Coding 37
4.2.2.1 Temporal Scalability 38
4.2.2.2 Spatial Scalability 38
4.2.2.3 Quality Scalability 40
4.3 Rate Control in Scalable Video Coding (SVC) 40
4.3.1 The Importance of Rate Control 41
4.3.2 The Rate-Quantization Model 42
4.3.3 Complexity Estimation 42
4.3.4 Rate Control Scheme 43
4.4 Related Rate Control Algorithm in SVC 44
4.5 Proposed Algorithm 45
4.6 Experimental Results 50
4.7 Summary 53

CHAPTER 5 54
INTER-VIEW REFERENCE FRAME SELECTION IN H.264/MVC 54
5.1 Introduction 54
5.2 H.264/Multiview Video Coding 55
5.2.1The MVC Applications: 55

5.2.2 The Requirements for MVC 55
5.2.3 Prediction Structure 56
5.2.3.1 Temporal Prediction Using Hierarchical B Pictures 57
5.2.3.2 Inter-view Prediction for Key Pictures 60
5.2.3.3 Inter-view Prediction for Key and Non-key Pictures 61
5.3 Related Efficient Algorithm for Speeding up the Prediction Process 62
5.4 Proposed Algorithm 63
5.5 Experimental Results 68
5.6 Summary 72

CHAPTER 6 73
CONCLUSION 73
6.1 Introduction 73
6.2 Main Contributions and Results 74
6.2.1 A Simplified Rate Control Algorithm for H.264/SVC 74
6.2.2 Inter-view Reference Frame Selection in H.264/MVC 75
6.3 Future Work 76
6.4 Summary 78
References 80

APPENDIX A 86
HIGH-COMPLEXITY MOTION ESTIMATION AND MODE DECISION IN H.264/AVC 86
A.1 Overview 86
A.2 The Lagrangian Cost 86

APPENDIX B 89
THE H.264 TRANSFORM, QUANTISATION, RESCALING AND INVERSE TRANSFORM PROCESS 89
B.1 4 × 4 Residual Transform and Quantisation in H.264: 89
B.2 4 × 4 Luma DC Coefficient Transform and Quantisation (16 × 16 Intra-mode) 93
B.3 2 × 2 Chroma DC Coefficient Transform and Quantisation 94

APPENDIX C 96
POST-ENCODING STAGE OF RATE CONTROL IN SCALABLE VIDEO CODING 96

APPENDIX D 99
PUBLICATIONS 99

APPENDIX E 111
CD-ROM IN THE BACK POCKET 111
E.1 Scalable Video Coding (SVC) - Experimental Results. 111
E.2 Multiview Video Coding (MVC) - Experimental Results. 111
E.3 Ph.D Oral Defense - Presentation 111

List of Figures

Figure 2.1 - Video compression process 10
Figure 2.2 - Overview of the encoder and decoder 11
Figure 2.3 - The prediction residual created process 11
Figure 2.4 - A group of pictures in display order 12
Figure 2.5 - A slice structure 13
Figure 2.6 - The structure of macroblock 14
Figure 2.7 - Three sample format patterns 15
Figure 2.8 - Macroblock partitions 16
Figure 2.9 - Hadamard transform matrix 17
Figure 2.10 - The prediction samples for a 4×4 block 18
Figure 2.11 - Eight prediction directions for the intra 4×4 block 19
Figure 2.12 - Mode 0, Mode 6 and Mode 2 intra prediction modes 19
Figure 2.13 - Three different transform matrix in H.264/AVC 20
Figure 2.14 - The zig-zag scan order 22
Figure 3.15 - Motion estimation search range 25
Figure 3.16 - Diamond search pattern 29
Figure 3.17 - The reference image used for motion estimation 31
Figure 3.18 - The current image 32
Figure 3.19 - The full search method result 32
Figure 3.20 - The four step search method result 33
Figure 3.21 - The diamond search method result 33
Figure 3.22 - The motion vector field by using phase correlation method 34
Figure 3.23 - The phase correlation method result 34
Figure 4.24 - Hierarchical B-pictures prediction structures for temporal scalability 38
Figure 4.25 - Scalable spatial coding procedure 39

Figure 4.26 [37] - Multilayer structure with additional inter-layer prediction for enabling spatial scalable coding 40

Figure 4.27 - The relationship between Bitrate and QP 41

Figure 4.28 - The relationship between predict MAD, actual MAD and QP (Foreman) 46

Figure 4.29 - The relationship between predict MAD, actual MAD and QP (Crew) 46

Figure 4.30 - Block diagram of the proposed rate control scheme at frame level 47

Figure 4.31 - Comparison between the standard and proposed reconstructed frame under same condition 51

Figure 4.32 - Comparison between the standard and proposed reconstructed frame under same target bitrate 53

Figure 5.33 - Hierarchical coding structure for temporal prediction with GOP 8 57

Figure 5.34 - Basic structures for coding with GOP 12 58

Figure 5.35 - Basic structures for coding with GOP 15 58

Figure 5.36 - Temporal prediction using hierarchical B pictures in the simulcast multi-view video coding 59

Figure 5.37 - Inter-view prediction for key pictures 60

Figure 5.38 - Inter-view prediction for key and non-key pictures 61

Figure 5.39 - The phase correlation between view1 and view0, view2 64

Figure 5.40 - The cropped parts of inter-view reference and current frame 66

Figure 5.41 - The flow chart of the inter-view reference frame skip decision 67

List of Tables

Table 3.1 - Mean opinion score (MOS) 27

Table 3.2 - The performance of the motion estimation search method 35

Table 4.3 - Test conditions 50

Table 4.4 - Comparison between the proposed algorithm and the JSVM 9.19.9 software 51

Table 4.5 - Test conditions 52

Table 4.6 - Comparison between the proposed algorithm and the JSVM 9.19.9 software with the same bitrate 52

Table 5.7 - The phase correlation test configuration 65

Table 5.8 - The phase correlation test results 65

Table 5.9 - The MVC test video sequences 68

Table 5.10 - The configurations setting 68

Table 5.11 - The view coding order relationship in MVC 69

Table 5.12 - The Ballroom result 70

Table 5.13 - The Exit result 70

Table 5.14 - The Race1 result 71

Table 5.15 - Performance comparison between the proposed method and standard 71

Acknowledgements

I would like to give my respect and thanks to my supervisor team: Prof. Djamel Ait-Boudaoud, Dr. Martin Varley, Dr. Stephen Mein and Dr. Abdelrahman Abdelazim, for their guidance and advices in the area of video compression throughout my research work.

I would like to thank my research degree tutor Dr. Bogdan Matuszewski for his support on registration and transfer during my research study.

I also wish to thank the staff of graduate research school office for their useful information and arrangement on the research skill training courses. Also I express my thanks to the School of CEPS and University Travel Office for their arrangement when I participated in international conference.

I would like to thank you for interesting in my thesis and appreciate my work.

Finally, I would like to thank my parents and my family for their constant support and encouragement during the past four years.

List of Abbreviations

AVC	Advanced Video Coding
CABAC	Context-based Adaptive Binary Arithmetic Coding
CAVLC	Context-based Adaptive Variable Length Coding
CCITT	International Telegraph and Telephone Consultative Committee
DCT	Discrete Cosine Transform
DPCM	Differential Pulse Code Modulation
FSBM	Full Search Block Matching
FTV	Free Viewpoint Television
GOP	Group of Picture
HD	High Definition
HEVC	High Efficiency Video Coding
HVS	Human Visual System
IDR	Instantaneous Decoder Refresh
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
ISO/IEC JTC 1	Joint Technical Committee 1 of the ISO and IEC
ITU-T	Telecommunications Union-Telecommunication Standardization Sector
JCT-VC	Joint Project Between VCEG and MPEG
JM	Joint Model
JPEG	Joint Photographic Experts Group
JSVM	Joint Scalable Video Model
JVT	Joint Video Team
MAD	Mean Absolute Distortion
MB	Macroblock
MC	Motion Compensation
ME	Motion Estimation
MPEG	Motion Picture Experts Group
MSE	Mean Square Error
MVC	Multiview Video Coding
PAL	Phase Alternating Line
PSNR	Peak Signal to Noise Ratio
QP	Quantisation Parameter
RD	Quantisation Parameter
RDO	Rate Distortion Optimization
SAD	Sum of Absolute Difference
SAE	Sum of Absolute Error
SATD	Sum of Absolute Transformed Differences
SSD	Sum of the Squared Differences
3DTV	Three-dimensional TV
UVLC	Universal Variable Length Coding
VCEG	Video Coding Experts Group
VHS	Video Home System

CHAPTER 1

INTRODUCTION

1.1 Motivation

Digital images and videos have been used widely in current communication networks, which include digital television, digital cameras, internet video and video conferencing. As a consequence, the representation of the information requires a very large amount of data. In addition, high definition television videos require more bits to be coded. For example, if a video sequence with a frame size of 720×480 is transmitted through a network at a rate of 25 frames/second, then about 207.4 million ($=720 \times 480 \times 8 \times 3 \times 25$) bits per second (bit/s or bps) of bandwidth is required. Similarly, if a high definition video sequence with frame size 1920×1080 is transmitted at a rate of 60 frames/second, about 3.0 Gbit/s of bandwidth is required. However, the common internet connection bandwidth is from 1.5 Mbit/s to 10 Mbit/s, which is too low to meet the requirements of modern consumer applications and devices. With the limited capabilities of transmission networks, more advanced and innovative compression techniques have been developed to meet the users' demands.

Compression coding techniques reduce the amount of data, thereby enabling the storage capacities and bandwidth of the transmission network to be sufficient for the compressed video data. However, many novel communication applications are emerging rapidly in recent years, such as three-dimensional television, movies on demand, games, smart phone, and surveillance equipment. These various applications are required to be transmitted over different bandwidths and storage capacities. Compression techniques play an important role in these applications; consequently, the scalable and multiview video coding techniques are proposed to solve these problems. The task of scalable video coding is to generate a single video stream that can be self-decoded to adapt to variety of devices with different capabilities. Multiview video coding is a fundamental technique for the three-dimensional video coding, in which a number of cameras are used to capture the same scene at the same time from different locations. Since the multiple video sequences contain a large amount of data, multiview video coding techniques are proposed to reduce the number of bits for efficient transmission and storage.

The H.264/AVC video coding standard can achieve the best compression with high image quality and reduced bitrate, but the computational complexity is significantly increased. The trade-off relationship between the computational complexity and performance (distortion and bitrate) is a major challenge in video coding optimization design. Algorithms with low computational complexity, but without significant loss in image quality or increase in bitrate, are the main research target in this thesis, As a consequence, applications can be used in the computational constraint environment, such as hand-held devices, high definition television and three dimensional games.

1.2 History of Video Compression

Many video coding standards have been developed and published by the international standards development organizations, which include the International Organization for Standardization (ISO) [1], the International Electrotechnical Commission (IEC) [2], and the International Telecommunications Union-Telecommunication standardization sector (ITU-T) [3]. The Video Coding Experts Group (VCEG) is a working group of the ITU-T, which was formed in 1984 and was responsible for the H.26x video coding standards. ISO/IEC JTC 1 is Joint Technical Committee 1 of the ISO and IEC, which was formed in 1987 [4]. The Moving Picture Experts Group (MPEG) was formed in 1988, and is a working group of experts from ISO/IEC aimed at developing standards for digital audio and video coding and transmission [5]. In order to deal with the overlap in areas of standardization and meet the global business and consumer requirements, the standardisation organisations often work in collaboration with each other. The Joint Video Team (JVT) and the Joint Collaborative Team on Video Coding (JCT-VC) are joint projects between VCEG and MPEG. The JVT was formed in 2001 and the H.264/MPEG-4 AVC was completed in 2006. The JCT-VC was formed in 2010 and the H.265/MPEG-H Part 2 High Efficiency Video Coding (HEVC) [6] was completed in January 2013. A brief history of video coding standards is described in the following section.

The first digital video coding standard H.120 was published by the International Telegraph and Telephone Consultative Committee (CCITT) in 1984 [7]. The CCITT was renamed ITU-T in 1993. The standard is based on differential pulse-code modulation (DPCM) techniques. Each sample or pixel is predicted from the previously

coded samples. The prediction samples can be the adjacent pixels within the same frame or in the previous frame, and the prediction error is transmitted and encoded. The H.120 standard can achieve a very good spatial resolution, but the temporal quality is very poor. The standard incorporates the Discrete Cosine Transform (DCT), zig-zag scan, variable-length coding and scalar quantization techniques. H.120 is designed to work at a bitrate of 1544 kbit/s for National Television System Committee (NTSC) and 2048 kbit/s for Phase Alternating Line (PAL). The standard was revised in 1988, and the motion compensation and background prediction were introduced in the standard.

The first successful practical digital video compression standard, H.261, was developed by ITU-T in 1900 and ratified in 1988 [8]. H.261 is the basis for many video coding standards. The new technical features used in the H.261 standard include 16×16 macroblocks, motion compensation, and run-level variable-length code. H.261 is designed to work at bitrate between 64 kbit/s and 2 Mbit/s. The standard was revised in 1993 and a backward-compatible mode for sending still images is supported.

MPEG-1 [9] was developed by ISO/IEC JTC 1 in 1993. The MPEG-1 standard was designed to achieve Video Home System (VHS) video quality at bitrate of 1.5 Mbit/s. It has been used in a large number of products, especially on video CD (VCD). MPEG-1 was developed based on H.261, and some new features were added such as bi-directional motion prediction, half-pixel motion compensation, D-type picture, and quantization weighting matrices. However, it does not support interlaced-scan pictures, while the NTSC and PAL video formats are interlaced.

The MPEG-2/H.262 [10] standard was developed jointly by the ITU-T VCEG and ISO/IEC MPEG in 1994. The standard was designed to be usef for DVD, standard and high definition television at a higher bitrate, and it outperforms MPEG-1 at bitrates of 3 Mbit/s or above. It also provides support for interlaced video. MPEG-2 is compatible with MPEG-1, and additional functions such as signal-to-noise scalable and spatial scalable are added to support the multiple resolutions for both the standard and high definition television. MPEG-3 was developed and targeted at high definition television, but it was replaced by MPEG-2 with the same performance.

The H.263 [11] standard was developed by the ITU-T VCEG in 1995. It was based on H.261, MPEG-1 and MPEG-2, and aimed to operate at a low bitrate for video conferencing. H.263 can provide higher quality at all bitrates in comparison with the prior standards. The original standard was modified in 1998 and 2000, and the coding

efficiency and capacities were significantly improved. A large number of new features were developed, for example, reference picture selection mode, modified quantization mode and support for flexible picture formats.

The MPEG-4 [12] visual standard was developed by ISO/IEC MPEG in 1998 and originally aimed at low bitrate video communications. The standard was expanded later to work efficiently from low bitrate up to high bitrate. The coding efficiency is enhanced compared to MPEG-2; in addition, new features are added in the standard, such as error resilience and segmented coding of shapes. More features were added during the development in 2000 and 2001, for example, variable block size motion compensation, intra discrete cosine transform coefficient prediction and quarter-sample motion compensation.

H.264 / MPEG-4 part 10 [13] was developed jointly by the ITU-T VCEG and ISO/IEC MPEG, which is known as JVT. It was started in 2001 and the first version was completed in 2003. The coding efficiency is significantly improved with quality kept at the same level in comparison to prior video coding standards. It has been used in many latest video coding applications, such as YouTube, Adobe Flash Player and Blu-ray Discs. A number of new features are included in the standard, for example, the number of reference pictures is increased up to 16 frames, more flexible variable block-size motion compensation is used with block sizes as small as 4×4 and as large as 16×16, and quarter-pixel motion compensation is included. The two major new features added to the standard were Scalable Video Coding (SVC) and Multiview Video Coding (MVC). The scalable video coding extensions were completed in 2007 and the multiview video coding extensions were completed in 2009. The details are described in the chapter 4 and chapter 5 of this thesis respectively.

H.265/MPEG-H Part 2 [14] High Efficiency Video Coding is a successor to H.264 / MPEG-4 advanced video coding and aimed to significantly improve the compression performance in comparison to existing standards. The high efficiency video coding standard was completed in January 2013 and can achieve a bitrate reduction of 50% for equal perceptual video quality, and can support increased video resolution such as ultra high definition television.

The standardization efforts have been developed based on the prior standards to meet the requirements of modern devices and services. The overall video compression efficiency is increased with the development of the innovative coding tools and

optimized coding algorithms. The coding parameters and variations involved in the video compression are presented in the following section.

1.3 Concepts and Definitions

The quality of the decoded video sequence is dependent on the choice of the video compression standard. A well-design video compression standard can achieve the best compression performance by optimizing the relationship between bitrate, image quality and computational complexity. The coding parameters are determined by the practical constraints of transmission and processing, such as bandwidth and power limited devices. The coding parameters and definitions involved in the video coding design are described below:

Image resolution

The image resolution defines the number of sampling points in each image. The various media use different resolutions, for example, DVD uses a resolution of 720×480, high definition television uses a resolution of 1920×1080 and ultra high definition television uses a resolution of 7680×4320.

Frame rate

The number of frames displayed per second in the video sequence is termed the frame rate. A higher frame rate gives a smoother scene but requires more bits to encode. The frame rate is determined by the bandwidth and applications. For example, standard television operates at 25 frames per second, while high definition television operates at 60 frames per second.

Bitrate

Bitrate is the number of bits that are required to be stored or transmitted per unit time. The various standards are designed for particular applications with different bitrate requirements. The MPEG-1 standard is designed for an intermediate bitrate of 1.5 Mbit/s, MPEG-2/H.262 is intended for a higher bitrate (about 10 Mbit/s), MPEG-4 is primarily intended for a low bitrate and is expanded to work efficiently across a variety of bitrates, and H.264/MPEG-4 AVC is intended for a very broad application range from low bitrate internet streaming applications to high definition television broadcast

with lower bitrate in comparison to previous standards [15]. The higher bitrate requires more processing power to decode.

Processing power

The processing power is dependent on the bitrate, image quality and algorithm techniques used in the different standards. The more exhaustive computation typically costs more processing power.

Latency

Some applications require very low latency transmission, such as video conferencing, while streaming video can tolerate a higher latency. The latency is depending on the picture type and buffer size, the B picture can increase the coding efficiency at the expense of a great latency. The high bitrate may also increase the latency in the network.

Error resilience and robustness

Error resilience is important when errors may occur during transmission. The bitstream should still be decoded when errors occur. Intra frame coding can limit the effect of transmission errors with the increased bitrate. An increased number of reference frames can improve the error resilience at the expense of increased complexity and storage. Resynchronisation methods can also be used to limit the error propagation.

Synchronization

An I-picture is set as the synchronization point and is useful for error resilience and resynchronization, since it is coded without prediction and can be decoded independently. However, I-pictures incur the cost of a higher bitrate in comparison to P-pictures and B-pictures.

1.4 Research Objectives

A video coding algorithm can achieve higher visual quality for a given bitrate at the expense of increased computation. For example, full search motion estimation can reduce the residual data by searching every possible position to find the best match; in addition, the increased search area can further reduce the redundancy data and bitrate.

However, it is impractical in reality because the computational complexity is related to the processing power, especially in power-constrained environments such as mobile or hand-held computing platforms. Furthermore, the coding computational complexity is increased with the high frame rate in the high definition television.

The H.264/AVC standard can compress the video with lower bitrate while retaining the same picture quality compared to previous standards, however, the performance increased the overall coding complexity. The aim of this research is to reduce the computational complexity of H.264/AVC standard and its extension.

1.5 Research Outline and Contributions

The scalable video coding and multiview video coding are two important features that are incorporated in added to the H.264/AVC standard. Scalability allows the streams to be encoded once but decoded to meet the different requirements in frame rate and picture quality. Multiview video coding is the basic coding technique for the three dimensional video coding. The two parts of the coding techniques are important research area in the future, but the techniques have not yet been widely adopted by the industry. The major problem focuses on the increased computational complexity. The computational complexity optimization algorithms used in the scalable video coding and multiview video coding is briefly described below:

1) Scalable video coding enables a single bitstream encoded once with hierarchical layers, with each layer meeting different requirements in frame rate, quality and resolution. Rate control is exploited to efficiently control the coded bitrate in each layer according to the available bandwidth. Rate control can achieve the best rate-distortion performance in each layer, but at a cost of increased computational complexity. In order to reduce the coding complexity, a simplified rate control algorithm is proposed in this work, and the complexity is successfully reduced compared to the standard.

2) Multiview video coding is an efficient coding technique for coding the multiple views captured at the same time from different viewpoints. The statistical dependencies between the adjacent views are exploited to

reduce the huge amount of data. Inter-view reference frames are used to reduce the bitrate, however, the computational complexity is increased as the number of reference frames is increased. The phase correlation technique is utilized to abandon the unnecessary inter-view reference frame according to the relationship between the current frame and the inter-view reference frame.

1.6 Organisation of the Thesis

The chapters of the thesis are organized as follows:

Chapter 2 provides an overview of video compression. The fundamental coding tools and function of H.264/AVC are described in detail.

Chapter 3 discusses the motion estimation search strategy in H.264/AVC and a comparison between full search motion estimation and several fast search methods is presented.

Chapter 4 presents a discussion on the scalability and rate control scheme of the H.264 scalable video coding standard. A simplified rate control algorithm in the scalable video coding is proposed and presented in this chapter. The computational complexity is successfully reduced in comparison to the standard.

Chapter 5 introduces the multiview video coding techniques in the H.264 multi-view video coding standard. The inter-view reference frame prediction technique is described in detail in addition to the temporal-view reference frame. The phase correlation technique is utilized in the reference frame manager, and the inter-view reference frame is selectively disabled based on the phase correlation coefficients. With the number of reference frames reduced, the computational complexity is successfully reduced.

Chapter 6 presents a summary of the techniques and proposed methods in the video compression, highlighting the original contribution of this work. A brief future of video coding and ideas for further investigation in the area of video compression are introduced.

CHAPTER 2

AN OVERVIEW OF VIDEO CODING

2.1 Introduction

The aim of video coding is to produce a set of bits (the bitstream) that represents the captured digital video sequence. However, the produced digital video bitstream in its uncompressed raw form is a large amount of data, and it is therefore difficult to store and transmit. Video compression techniques have been developed to convert the huge data to a compressed format and enable the video to be reconstructed video with high quality. The compression is achieved by removing or reducing the redundancy in the temporal, spatial and/or frequency domain from the signal. A well designed video compression standard can provide a better image quality after reconstruction, leading to high reliability and flexibility on a variety of applications. With the development of the compression tools and compression algorithms, more and more advanced video compression standards have been proposed.

The latest H.264/AVC standard is an efficient and robust standard which supports different applications on storage, transmission and streaming. This chapter presents an overview of the H.264/AVC standard and basic concepts in video compression. The common video compression process is described in detail in the following section.

This chapter is organised as follows: An overview of video compression is presented in section 2.2. The fundamental coding tools and functions in H.264/MPEG Part 10 standard are described in section 2.3, and the summary is given in section 2.4.

2.2 Overview of Video Compression

Video compression is a process that represents the original digital video sequence in a smaller number of bits, and it consists of two parts: the encoder and the decoder. The encoder is used to convert the original video data into a reduced number of bits, and the decoder is used to recover the original video data from the reduced number of bits. The video compression process is depicted in figure 2.1.

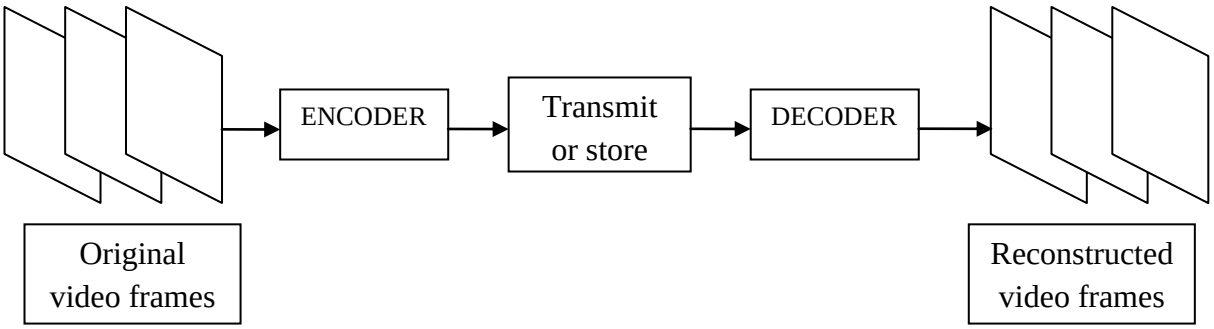


Figure 2.1 - Video compression process

Video compression techniques can be classified into two types: lossless compression and lossy compression. If the output reconstructed video data exactly matches the original video data, it is called lossless compression. With lossy compression, the reconstructed video data is only an approximation to the original video data. However, lossy compression techniques can achieve a higher compression rate, since the unnecessary element of the image is removed without significantly affecting the viewer's perception of visual quality.

The encoder process is much more complex than the decoder, since it needs to remove the redundancy between relative frames and produce a compressed bitstream before transmission or storage. In order to achieve a high compression efficiency and image quality, the encoder is required to determine the optimal motion vectors, quantization parameters and number of bits. The number of bits is constrained by the variety of application and bandwidth. Conversely, the decoder reconstructs the video frames from the compressed bitstream. The decoder must prevent the buffer from overflow or underflow by utilizing the time information and other parameters in the compressed bitstream. Figure 2.2 shows the overview of encoder and decoder process.

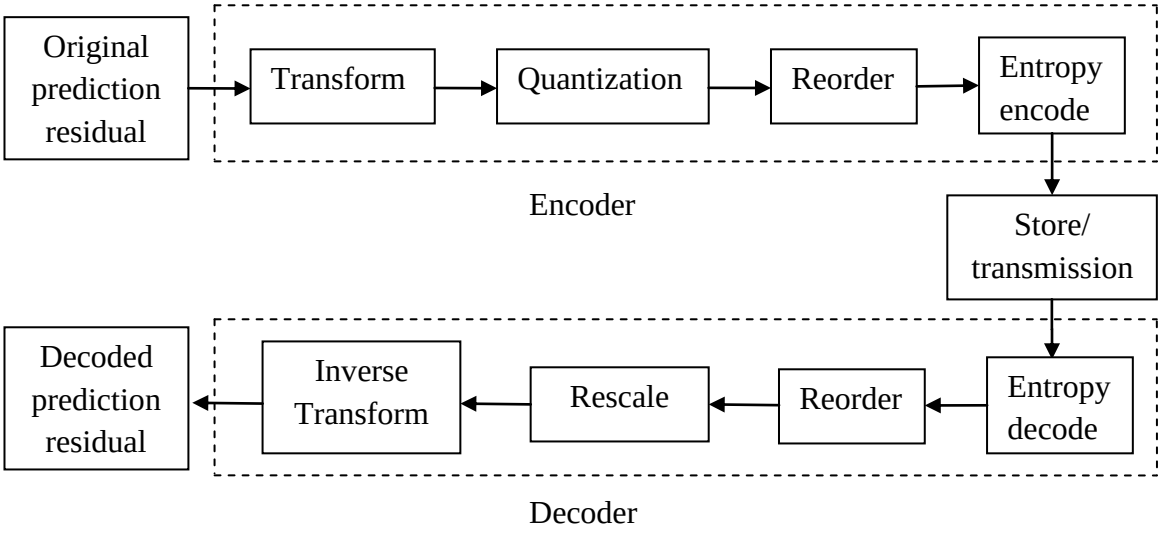


Figure 2.2 - Overview of the encoder and decoder

Intra- or inter-prediction techniques are used to produce the prediction residual. Intra-prediction techniques use prediction information only within the same frame, whilst inter-prediction techniques use prediction information from one or more previous or future coded frames in display order. The residual is created by subtracting the prediction from the current macroblock. Inter-frame prediction can achieve better compression performance by utilizing the similarities between adjacent frames, whereas intra-prediction performs better when the scene in adjacent frames is significantly different. Video compression techniques often combine the intra- and inter-prediction together to improve the compression efficiency and image quality. The residual creation process is depicted in figure 2.3.

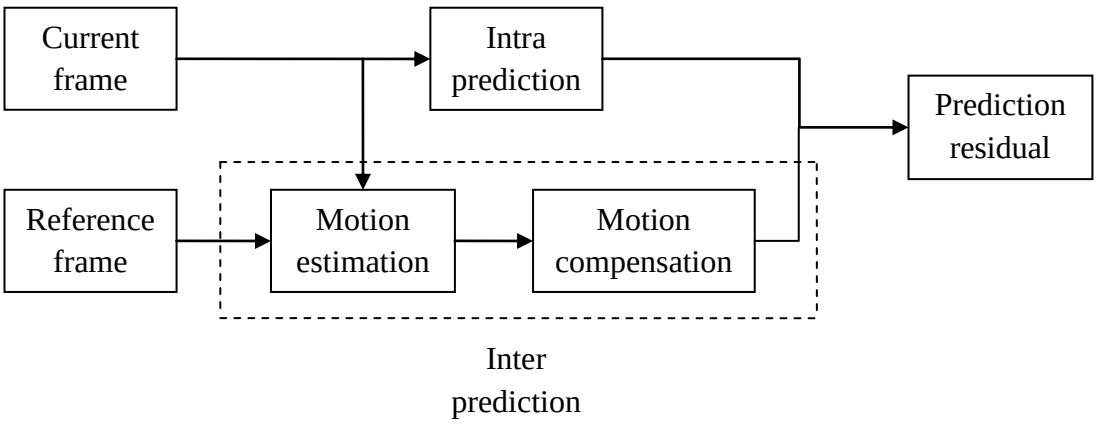


Figure 2.3 - The prediction residual created process

2.3 H.264/MPEG Part 10 Standard

2.3.1 Introduction

The latest advanced video coding standard is published as Part 10 of MPEG and ITU-T Recommendation H.264 [16]. The H.264 standard can provide better compression rate and video quality than previous standards. Since the basic functional elements such as transform, quantisation, entropy coding process are commonly used in video coding standards, the concept and definition in the H.264 standard is used to present the overall video compression process.

2.3.2 Concept and Definition

Video sequence

A video sequence is a series of pictures taken at a constant time intervals. As a consequence, video compression techniques take advantage of the similarities between adjacent pictures to reduce the bitrate. Each video sequence is divided into one or more group of pictures, and each group is encoded with three different types: I-picture, P-picture, or B-Picture. A group of pictures in display order is depicted in figure 2.4.

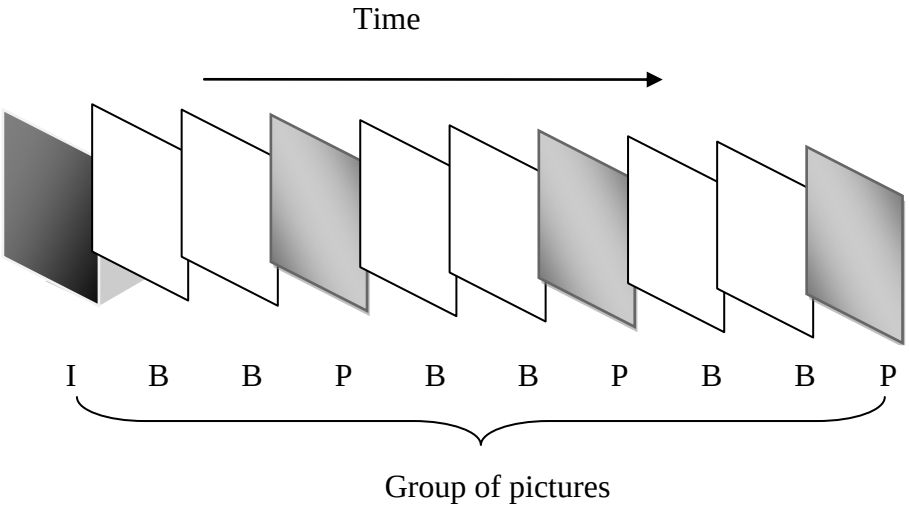


Figure 2.4 - A group of pictures in display order

I-pictures are intra coded pictures, which are encoded without reference to other pictures. Each macroblock in an I-picture is predicted from a previously coded macroblock within the same picture.

P-pictures are predictive coded pictures that are predicted from the preceding I-pictures or P-pictures. Each macroblock in a P-picture is predicted from the macroblock in preceding pictures or without prediction. If without prediction, the macroblock will be intra coded.

B-pictures are bidirectional predictive coded pictures that are predicted from the preceding or future I-pictures or P pictures. Each macroblock in a B-picture is predicted from preceding and/or future pictures, or intra coded.

Slice

A coded video picture is composed of a number of slices; each slice consists of a number of contiguous macroblocks in raster order. The number of macroblocks in each slice toned not be constant; it can be just one macroblock or the total number of macroblocks in a picture. The slice structure enables the encoder to have a great flexibility on controlling the coding parameters, especial in rate control. A slice structure is depicted in figure 2.5.

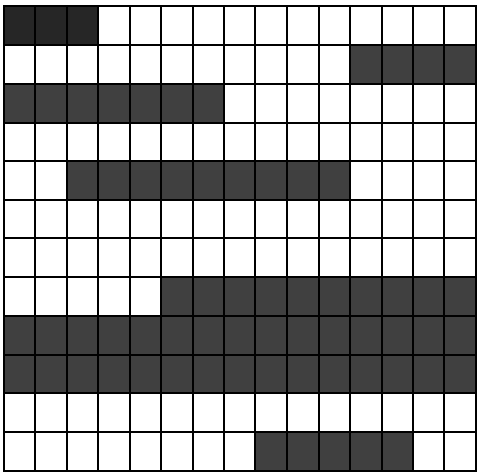


Figure 2.5 - A slice structure

Macroblock

Macroblock (MB) is the basic unit in a picture, and corresponds to a 16×16 sample region. Each region contains 16×16 luminance samples, 8×8 Cb and 8×8 Cr samples in a 4:2:0 format [17]. The 16×16 luminance sample is composed of four 8×8 blocks of

samples, and can be divided into more small regions for motion compensation. The structure of a macroblock in a 4:2:0 format is depicted in figure 2.6.

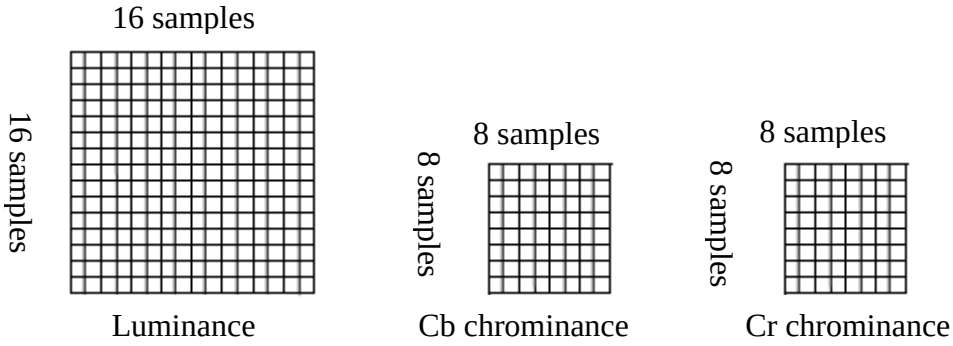


Figure 2.6 - The structure of macroblock

Colour Space

The colour space is used to represent the colour information in the video sequence. The RGB and YCbCr colour space are the most popular in video compression. In the RGB colour space, any colour sample can be represented by the three colours: Red (R), Green (G) and Blue (B). For example, mixing red and green produces yellow, and mixing red and blue produces magenta [18]. However, mixing all of the colours together cannot produce pure spectral colours.

In the YCbCr colour space, the colour sample is represented in luminance and chrominance. Since the human visual system is more sensitive to luminance than colour, the YCbCr colour space separates the luminance information from the colour information [19]. Therefore, it is more efficient to represent luminance with a higher resolution than colour. The luminance component Y is represented as a weighted combination of Red, Green and Blue [20]:

$$Y = k_r R + k_g G + k_b B \tag{2.1}$$

where k are weighting factors. The chrominance sample components are represented by the difference between R, G or B and the luminance component Y:

$$Cb = B - Y$$

$$Cr = R - Y \tag{2.2}$$

$$Cg = G - Y$$

Since Cb+Cr+Cg is a constant and the third component can always be calculated from the other two, then only luminance (Y) and blue (Cb) and red (Cr) chrominance components are required in the YCbCr colour space. Because the eye is less sensitive to the chrominance components, a lower spatial resolution is used to represent the chrominance components. The three sample formats are depicted in figure 2.7.

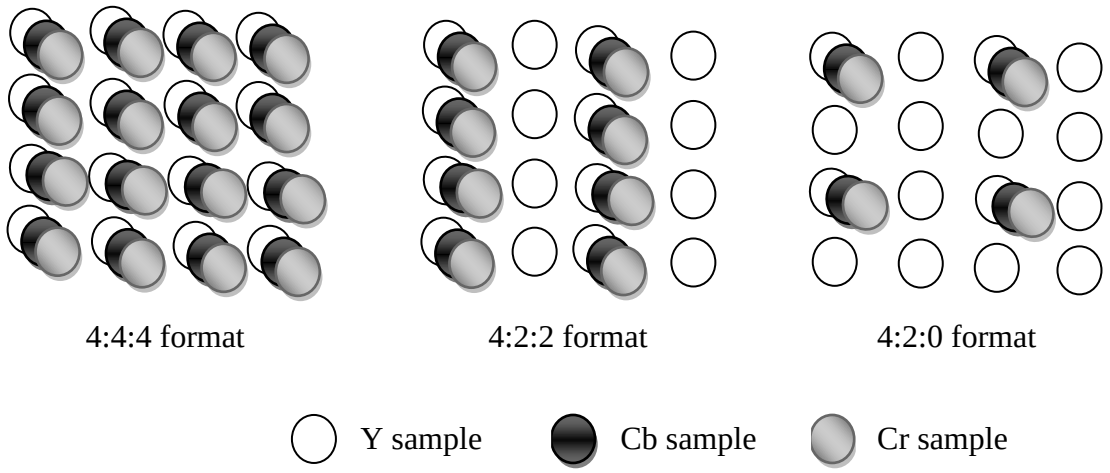


Figure 2.7 - Three sample format patterns

The 4:4:4 format indicates the full resolution of the three components. In 4:2:2 format, the vertical resolution of the chrominance is same as the luminance component, but the horizontal resolution of the chrominance component is reduced by a factor of 2. In 4:2:0 format, the spatial resolution of the chrominance is reduced by a factor of 2 in both the horizontal and vertical directions. While the spatial resolution of chrominance component is reduced, the number of bits is reduced.

2.3.3 Inter Prediction

Inter prediction is an important and efficient techniques used to reduce the input residual before transform. Inter prediction utilizes the similarities between adjacent frames to reduce the residual. The previously encoded video frames are regarded as the reference frames. The block based motion estimation method is used to find the best match in the reference frame for the macroblock in the current frame. The best match block in the reference frame is subtracted from the current block to produce the residual, and this step is described as motion compensation. The motion compensated stage is carried out on 16×16 macroblocks down to 4×4 blocks. The luminance component of each 16×16 macroblock may be split up in four ways: 16×16, 16×8, 8×16, and 8×8 macroblock partitions. If the 8×8 mode is chosen, each of the 8×8 partitions can be split into a further 4 ways: 8×8, 8×4, 4×8, 4×4 sub-macroblock partitions [21]. The macroblock partitions are shown in figure 2.8.

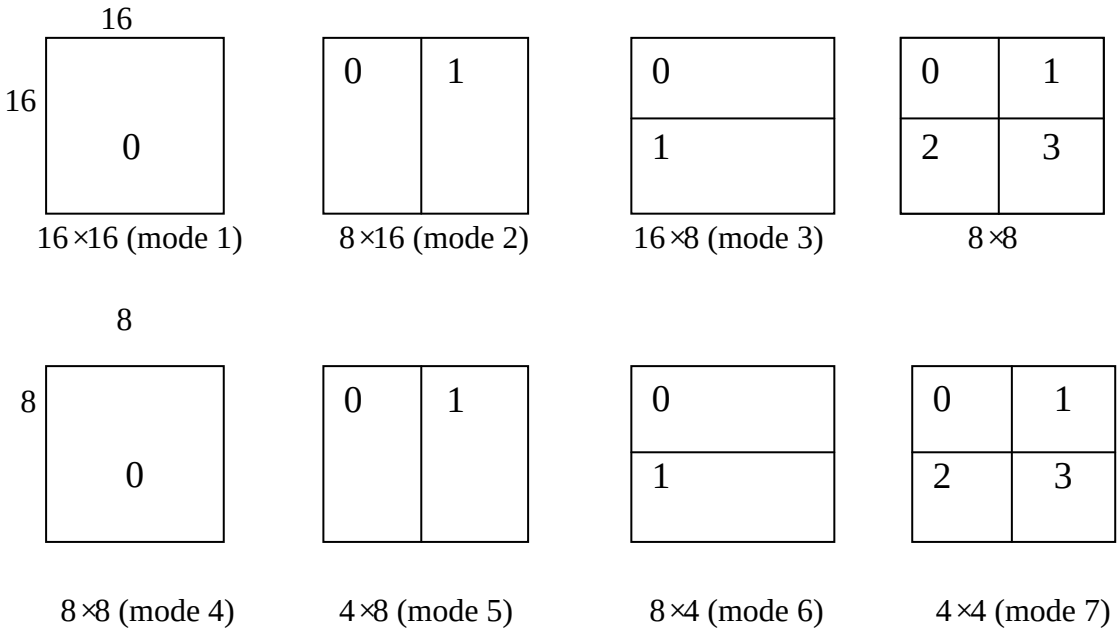


Figure 2.8 - Macroblock partitions

The small partition size can be more efficient at reducing the energy in the motion compensation, but it requires more bits to encode the motion vectors and the complexity is increased on the choice of partitions. The large partition size requires a small number of bits, but a significant amount of energy may contain in the motion compensation residual. In order to choose a best mode, the Lagrangian cost function [22] is used to

compute the cost for each mode and the mode that gives the smallest cost is selected. The Lagrangian cost function is defined as:

$$J = Distortion + \lambda_{MODE} \times Rate \tag{2.3}$$

The distortion indicates the quality of the reconstructed pictures and the energy remaining in the difference block, which can be computed using Sum of Absolute Differences (SAD), Sum of Absolute Transformed Differences (SATD), or Sum of Squared Differences (SSD). The SAD is computed using the following equation:

$$SAD = \sum_{ij} |S_{ij} - R_{ij}| \tag{2.4}$$

where S_{ij} represents the current block and R_{ij} represent the candidate prediction block in the reference frame. The SATD can achieve a better estimation of the distortion by estimating the effect of the discrete cosine transform (DCT) with the 4×4 Hadamard transform. The transform matrix is shown in figure 2.9.

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}$$

Figure 2.9 - Hadamard transform matrix

Since H is a symmetric matrix, it is equal to its own transpose. The SATD is computed using the following equation:

$$SATD = \frac{\sum_{ij} \left| H \times (S_{ij} - R_{ij}) \times H \right|}{2} \tag{2.5}$$

The sum of squared differences (SSD) between the current block and its reconstruction gives the real distortion, which is computed using the following equation:

$$SSD = \sum_{ij} \left[S_{ij} - C_{ij} \right]^2 \tag{2.6}$$

where C_{ij} represent the reconstructed block after decoding.

The Rate in Equation 2.3 indicates the total bits that used to code the macroblock using the particular mode. λ_{MODE} is the Lagrangian multiplier that used to quantify the

distortion and mode, which is decided based on the picture type. The mode that gives the smallest J is chosen as the best mode. The detail of the mode decision process in H.264/AVC standard is described in Appendix A.

Although the full exhaustive search in motion estimation increases the quality of the reconstructed video, the computational complexity increases significantly, since the Lagrangian cost function is computed for all possible modes. In order to speed up the coding process, several fast motion estimation techniques have been proposed and the details are described in chapter 3. Inter prediction can achieve a better compression performance when the successive frames are very similar; however, it is more efficient to use intra prediction when the scene is significantly different.

2.3.4 Intra Prediction

In the intra prediction process, the prediction block is predicted from the neighbouring, previously encoded blocks within the same frame. The residual is formed by subtracting the prediction from the current block. There are two intra prediction sizes for luminance samples in the H.264 standard: 4×4 block and 16×16 macroblock. The prediction samples for a 4×4 block are depicted in figure 2.10.

M	A	B	C	D	E	F	G	H
I	a	b	c	d				
J	e	f	g	h				
K	i	j	k	l				
L	m	n	o	p				

Figure 2.10 - The prediction samples for a 4×4 block

All samples of the current 4×4 block (a, b… p) are predicted from the neighbouring left or top samples (A, B...M), which are already encoded and reconstructed. There are nine prediction modes for 4×4 luminance block, eight prediction modes each for a specific direction and one DC prediction mode. The eight prediction directions are depicted in figure 2.11.

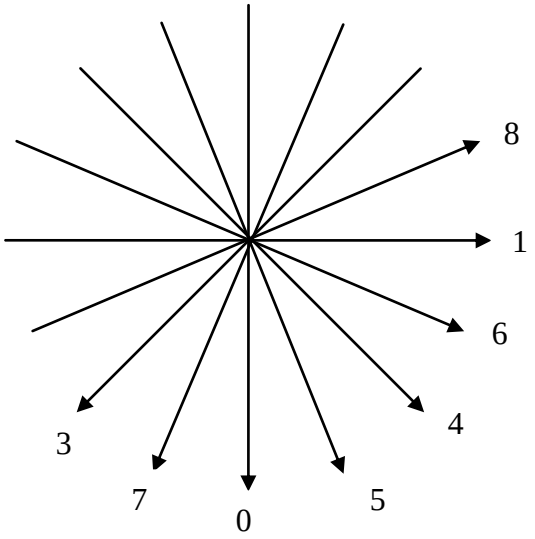


Figure 2.11 - Eight prediction directions for the intra 4×4 block

Mode 0, Mode 6 and Mode 2 are shown explicitly in figure 2.12.

M	A	B	C	D	E	F	G	H
I	a	b	c	d				
J	e	f	g	h				
K	i	j	k	l				
L	m	n	o	p				

Mode 0: Vertical

M	A	B	C	D	E	F	G	H
I	a	b	c	d				
J	e	f	g	h				
K	i	j	k	l				
L	m	n	o	p				

Mode 6: Horizontal-Down

M	A	B	C	D
I	a	b	c	d
J	e	f	g	h
K	i	j	k	l
L	m	n	o	p

Mode 2: DC (All samples are predicted by the mean of samples A...D and I…L).

Figure 2.12 - Mode 0, Mode 6 and Mode 2 intra prediction modes

For example, if vertical prediction (Mode 0) is chosen, then all samples below A are predicted from sample A. For mode 3-8, all samples are predicted from the weight average of the samples A-M. If mode 2 is chosen, all samples are predicted by the mean of samples A...D and I…L.

Four prediction modes are supported for the intra 16×16 block: Vertical prediction, horizontal prediction, DC-prediction and plane-prediction. Plane-prediction uses a linear function of top and left samples to predict the current samples.

2.3.5 Transform Coding

The purpose of the transform is to convert the residual data into transform domain and represented by a set of transform coefficients. In the transform domain, the energy in the image is grouped into a small number of significant values, and the insignificant data is discarded to reduce the number of bits. The discrete cosine transform is widely used in the data compression. The discrete cosine transform works on the residual data to produce the transform coefficients by using the following equation:

$$Y = HXH^T \tag{2.7}$$

and the inverse discrete cosine transform is given by:

$$X = H^T YH \tag{2.8}$$

where X is a matrix of the residual samples, Y is a matrix of the transformed coefficients and H is an N×N transform matrix.

Three different types of integer transform are applied in the H.264 standard. The three types of transform matrix are shown in figure 2.13 [23].

$$H_1 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \qquad H_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \qquad H_3 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

Figure 2.13 - Three different transform matrix in H.264/AVC

The three transform matrices are selected depending on the type of the residual data. The H_1 transform matrix can be applied to both the luminance component and chrominance components regardless of whether inter of intra prediction is used. If the macroblock is predicted using the intra 16×16 mode, then the Hadamard transform matrix H_2 is applied in addition to the first one. The H_3 matrix is used for the transform

of the chrominance components. The transformation process of H.264/AVC is described in detail in Appendix B.

2.3.6 Quantisation Stage

In order to remove the insignificant data in the transform coefficients, a quantisation stage is used after forward transformation. All transform coefficients are quantized by a quantisation parameter (QP) or quantiser step (Qstep). A total of 52 (0-51) values of QP and Qstep are supported in the H.264 standard. Qstep doubles in size for every increment of 6 in QP. The performance in terms of compression and image quality depend on the quantisation level. Each of the transformed coefficients Y_{ij} is quantised by integer division [24]:

$$Z_{ij} = round\ (\frac{Y_{ij}}{Qstep})$$

$$Y'_{ij} = Z_{ij} \times Qstep \tag{2.9}$$

where Z_{ij} is the quantised coefficients, Y'_{ij} is the reconstructed transform coefficients by simple scaling the quantised coefficients by $Qstep$. Small-valued coefficients become zero after quantisation. A large quantisation parameter means a significant reduction of the transform coefficients and gives higher compression rate at the expense of increased distortion in the decoded image. A small quantisation parameter means most of the transform coefficients are kept, but the compression rate is low. If the quantisation parameter is chosen correctly, all of the significant coefficients will be retained and the insignificant coefficients will be removed. The decoder will reverse the quantisation stage to reconstruct the transform coefficient; however, the insignificant coefficients cannot be replaced and so quantisation is a lossy process. The wide range of quantisation parameter can be used to achieve the optimum balance between the bitrate and quality. However, the encoder is required to code the macroblock repeatedly before selecting the optimal quantisation parameter that minimises the size if the encoded data. In addition, a set of parameters are involved in this process, for example, prediction mode (Intra of Inter), mode decision, and motion vector. Many rate-distortion optimization algorithms have been proposed (such as Lagrangian optimization algorithm) for the encoder to achieve optimum performance.

High efficiency video coding (HEVC) uses the same uniform reconstruction quantisation scheme controlled by a quantisation parameter as in H.264/MPEG-4 AVC [6]. Quantisation scaling matrices are also supported for the various transform block sizes. The details are provided in Appendix B.

The quantisation stage is followed by the entropy coding, and the remained significant transform coefficients together with the header and motion vector information are entropy coded to form the final compressed bit stream of the video sequence.

2.3.7 Entropy Coding

The purpose of entropy coding is to represent the compressed data using as few bits as possible. Entropy coding is a lossless compression technique, and the compression is achieved by exploiting the distribution of the quantised transform coefficients. After transformation and quantisation, the residual data is converted to a few non-zero coefficients and a large number of zero coefficients. The non-zero coefficients are typically the significant data that are located around the top-left corner of the array; the zero value coefficients represente the insignificant data. In order to encode the large number of zero coefficients more efficiently, the transform coefficients are reordered to a one-dimensional array. A commonly used zig-zag scan order is depicted in figure 2.14.

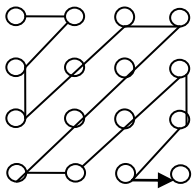


Figure 2.14 - The zig-zag scan order

The run-level technique is used to efficiently represent the large number of zero coefficients, where run indicates the number of zero coefficients preceding the non-zero and the level indicates the value of the non-zero coefficients. As a consequence, compression is achieved by representing frequently occurring symbols with a small number of bits and less frequent symbols with a large number of bits.

The two popular entropy coding methods used in the H.264 standard are context-based adaptive variable length coding (CAVLC) [25] and context-based adaptive binary arithmetic coding (CABAC) [26]. Context-based adaptive variable length coding takes advantage of the fact that non-zero coefficients in neighbouring blocks are related, the choice of the variable length coding look-up table for the level parameter depends on the previously coded level in the neighbour blocks. Context-based adaptive binary arithmetic coding is based on arithmetic coding; therefore, the transform coefficient is converted to binary first. A probability model is used for each binarized symbol, and the model is updated according to the recently coded value. Then, the arithmetic coder encodes each binarized symbol based on the probability model. The context-based adaptive binary arithmetic coding provides much better compression than context-based adaptive variable length coding, but it requires a larger amount of computation process.

2.4 Summary

The fundamental coding tools of H.264/AVC are introduced in this chapter, which include motion compensation, transform coding, quantisation and entropy coding. With the optimized video coding tools, the H.264 standard can achieve a higher compression efficiency compared to previous video coding standard. With the complex motion estimation and mode decision strategy, a higher image quality and impressive bitrate reduction is enabled in the H.264 standard. However, the best performance is achieved at the expense of the high computational complexity, this constrains the development of the small devices with limited computational resources and power consumption.

The computational performance, bitrate and video quality are major challenge in video compression, transmission and storage. The rate control algorithm is presented in chapter 4, which aims to achieve the best performance with the constraints of bitrate and complexity.

CHAPTER 3

MOTION ESTIMATION STUDY ON H.264/AVC

3.1 Introduction

Motion estimation and compensation play important roles in effective video coding standards, and the block-based motion estimation method is used to find out the optimal motion displacements between two frames. The Full Search [27] block matching motion estimation method used in the latest H.264/AVC standard is considered to be the most efficient motion estimation method; it evaluates the cost at each point in the search window to find the best match. On the other hand, the computational cost of the full search algorithm is very high. In order to overcome the constraints on computational complexity and power-limited applications, several fast search algorithms have been proposed, for instance, the Three Step Search, Logarithmic Search, Cross Search [27][28] [29]. However, these algorithms normally reduce the quality of the search and the motion displacement is not the optimal.

Besides the conventional block-matching methods, the block-based phase correlation technique is commonly utilised to estimate the translational displacement between two frames. The advantage of phase correlation method is lower computational complexity especially for large scale translation. A comparison of the Full Search, Four-step Search [30], Diamond Search [31] and phase correlation [32] methods is presented in this chapter.

This chapter is organised as follows: The motion estimation search strategy in H.264/AVC is presented in section 3.2. Several fast search motion estimation algorithms are presented in section 3.3. Section 3.4 presents the experimental results. Finally, the conclusion is given in section 3.5.

3.2 Motion Estimation in H.264/AVC

Motion estimation is an important step to reduce the residual between the current frame and reference frame in the encoder stage. The basic macroblock unit is a 16×16-pixel region, and the objective of motion estimation is to find a macroblock that closely matches the current macroblock in the reference to minimise the residual. In addition, the computational complexity is highly related to the size of the search range which is normally centred on the current macroblock in the reference frame. The motion estimation search range is depicted in figure 3.15. The full search algorithm calculates the cost at each point in the search area, the highest Peak Signal-to-Noise Ratio (PSNR) can be achieved through the sophisticated search method.

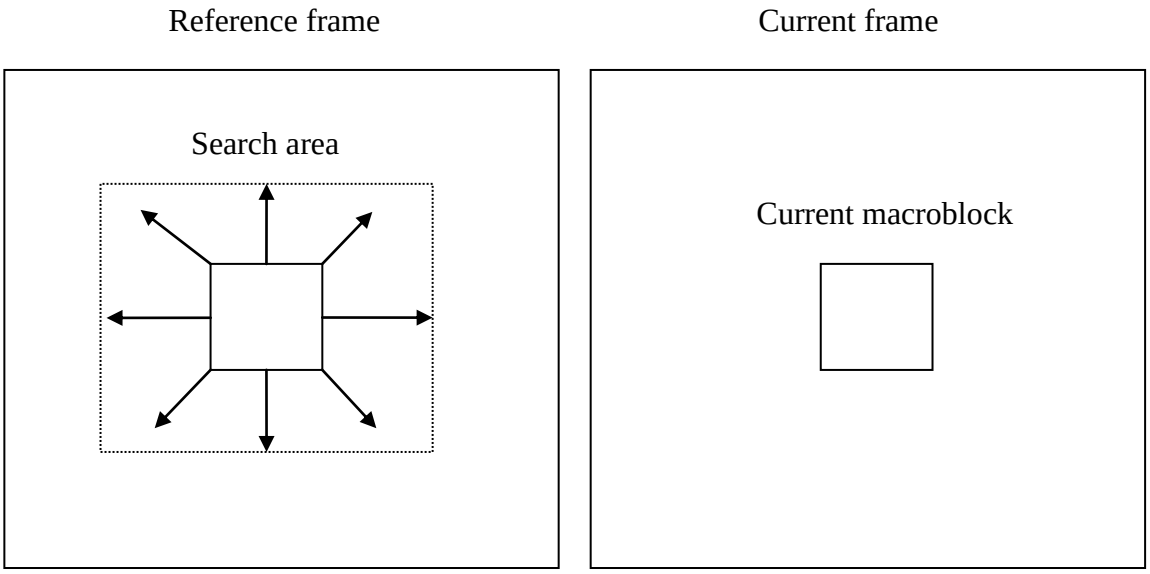


Figure 3.15 - Motion estimation search range

3.2.1 Criterion for Best Match

The best match of one macroblock with another is based on a cost function, such as Mean Square Error (MSE) (3.1), Mean Absolute Distortion (MAD) (3.2) and Sum of Absolute Error (SAE) (3.3) [27].

$$MSE = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\left(C_{ij} - R_{ij}\right)^2 \quad (3.1)$$

$$MAD = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\left|C_{ij} - R_{ij}\right| \tag{3.2}$$

$$SAE = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\left|C_{ij} - R_{ij}\right| \tag{3.3}$$

The N^2 represents the N×N samples in one macroblock, and C_{ij} and R_{ij} represent the current and reference macroblock respectively. The best match for the current macroblock in the reference frame is chosen as the one with the smallest cost.

3.2.2 Image Quality

Image quality measurement has been employed for quality control, which can be used to evaluate and obtain the best quality image and video data. Objective and subjective image quality assessments play an important role in various image processing applications. Image quality can be used to evaluate the performance of the various applications, and optimize the algorithms and parameters setting in these applications.

Peak signal-to-noise ratio (PSNR) (3.4) is used to measure the objective quality of the reconstructed image, and the mean square error (MSE) (3.5) is also involved to measure the distortion between the current frame and the reconstructed frame. A high PSNR value indicates a reconstructed frame with high quality, and conversely a low PSNR value indicates low quality.

$$PSNR_{dB} = 10\log_{10}\frac{(2^n-1)^2}{MSE} \tag{3.4}$$

$$MSE = \frac{1}{M \times N}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(C_{ij} - R_{ij}\right)^2 \tag{3.5}$$

where n represent the number of bits in each image sample, M×N indicate the image size, C_{ij} and R_{ij} represent the current and reconstructed frame respectively.

Objective assessment can measure the image and video quality automatically and quickly, however, it does not necessarily correspond well with perceived quality measurement. The subjective quality measurement Mean Opinion Score (MOS) has

been used for many years to obtain the human user's subjective opinion of the audio or video quality [33]. The mean opinion score is the arithmetic mean of all the individual scores, using the following rating scheme:

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Table 3.1 - Mean opinion score (MOS)

Two main methods for the subjective assessment of the quality of television systems are given in [34]. The double-stimulus impairment scale (DSIS) method is used to measure the robustness of systems, that is, failure characteristics. The double-stimulus continuous quality-scale (DSCQS) method is used to measure the quality of systems relative to a reference. Different environments with different viewing conditions are used for subjective assessment, which include the laboratory viewing environment and the home environment. In addition, a group of observers are needed in these assessments. Although subjective assessment is the best way to assess the quality of the reconstructed video by the ultimate receivers (humans) in real environments, it is too slow and expensive for practical usage. In practice, the imperfect distortion models such as the mean square error (MSE), sum of squared differences (SSD) or peak signal-to-noise ratio (PSNR) are usually used in performance comparisons.

3.3 Fast Search Algorithms

Several fast search algorithms are proposed to reduce the computational complexity, in comparison to the full search method, the search range is decreased and less search points are used. On the other hand, the quality of the reconstructed picture is reduced. The four-step search, diamond search and phase correlation fast algorithms are described and evaluated in this section.

3.3.1 Four-step Search Method

A novel four-step search algorithm is proposed in [29] with the centre biased search scheme. A fixed search step S=2 is set in this algorithm. The search procedure is described as follows:

1) First, nine locations are searched around the current macroblock in the reference frame, as depicted below:

x	-2	0	2	-2	0	2	-2	0	2
y	-2	-2	-2	-0	0	0	2	2	2

   If the least cost is found at the centre of the search area, then the search jumps to the fourth step, that is, the search step is reduced to S=1. Otherwise, the search move to the second step and the least cost position is set as the new origin.

2) The second step is same as the first step with new search origin, however, the search points are reduced, that is, only three or five new search points are required to be tested. For example, if the least cost is found at the (-2,0) or (2,0), only three new positions need to be checked. Taking the (-2,0) position as an example, the new search position is depecited below:

x	-4	-2	0	-4	-2	0	-4	-2	0
y	-2	-2	-2	-0	0	0	2	2	2

   It is only necessary to check the (-4, -2), (-4, 0), and (-4, 2) positions again to find the least cost. If the least cost is found at the centre of the search area, then the search jumps to the fourth step. Otherwise, the search moves to the third step and the least cost position is set as the new origin.

3) The second step is repeated in the third step.

4) In the fourth step, the step size is reduced to S=1. The position with the least cost is chosen as the best match.

3.3.2 Diamond Search Method

A diamond fast search motion estimation method is proposed in [31]. A diamond search pattern is used instead of the square pattern, and a large diamond search pattern and a small diamond search pattern are introduced in the [31]. The diamond search pattern is depicted in figure 3.16.

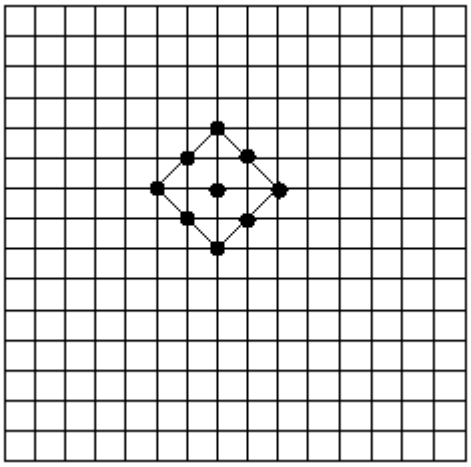


Figure 3.16 - Diamond search pattern

The diamond search algorithm is described as follows:

1) First, the initial nine locations are searched around the origin point by using the large diamond search pattern. If the least cost is found at the centre of the search area, the search jumps to the third step, and a small diamond search pattern is used, otherwise, it jumps to the second step.

2) The least cost point in the first step is set as the new origin, if the least cost is found at the centre of the search area, then go to step 3, otherwise, this step is repeated until the least cost is found at the centre of the search area.

3) A small diamond search pattern is used in this step. The least cost point in the second step is set as the centre of the search area, and the least cost point found in this step is regarded as the best match.

3.3.3 Phase Correlation Method

The phase correlation algorithm takes the inverse Fourier transform of the phase of the Fourier cross-power spectrum of a pair of image to extract the relative displacement vector. As well as the low computational complexity, the phase correlation algorithm is resilient to scene content, illumination differences and narrow-band noise [32]. However, the phase correlation algorithm only performs well with regard to a single moving object in the frame. In order to solve this problem, the reference frame and current frame are divided into small macroblock, and the phase correlation is calculated between each pair of co-located rectangular blocks. The optimal displacement vector is chosen with the maximum phase of the cross-power spectrum in each block. The basic principles are described as follows:

Assuming a translational shift between the two frames

$$s_{k+1}(n_1, n_2) = s_k(n_1 + d_1, n_2 + d_2) \tag{3.6}$$

Their discrete two-dimensional Fourier transforms are

$$s_{k+1}(f_1, f_2) = s_k(f_1, f_2)\mathrm{e}^{[j2\pi(d_1 f_1 + d_2 f_2)]} \tag{3.7}$$

Therefore the shift in the spatial-domain is reflected as a phase change in the spectrum domain. The cross-correlation between the two frames is

$$c_{k,k+1}(n_1, n_2) = s_{k+1}(n_1, n_2) \ * \mathrm{s}_\mathrm{k}^*(-\mathrm{n}_1, -\mathrm{n}_2) \tag{3.8}$$

whose Fourier transform is

$$C_{k,k+1}(f_1, f_2) = s_{k+1}(f_1, f_2).\mathrm{s}_\mathrm{k}^*(f_1, f_2) \tag{3.9}$$

The phase is obtained by normalizing the cross-power spectrum by its magnitude

$$\Phi\left[\mathrm{C}_{\mathrm{k,k+1}}(\mathrm{f}_1, \mathrm{f}_2)\right] = \frac{\mathrm{s}_{\mathrm{k+1}}^*(\mathrm{f}_1,\mathrm{f}_2).\mathrm{s}_\mathrm{k}(\mathrm{f}_1,\mathrm{f}_2)}{|\mathrm{s}_{\mathrm{k+1}}^*(\mathrm{f}_1,\mathrm{f}_2).\mathrm{s}_\mathrm{k}(\mathrm{f}_1,\mathrm{f}_2)|} \tag{3.10}$$

By equation (3.7) and (3.10), we have

$$\Phi\left[\mathrm{C}_{\mathrm{k,k+1}}(\mathrm{f}_1, \mathrm{f}_2)\right] = \mathrm{e}^{[-j2\pi(d_1 f_1 + d_2 f_2)]} \tag{3.11}$$

The two-dimensional inverse transform is given by

$$c_{k,k+1}(n_1, n_2) = \delta(n_1 - d_1, n_2 - d_2) \tag{3.12}$$

Then the motion vector is obtained by using the location of the pulse equation (3.12). The pulse is located by finding the highest peak.

3.4 Experimental Results

The above full search and several fast search methods are tested by using two successive frames in the 'Caltrain' video sequence. The reference image and the current image are shown in figure 3.17 and figure 3.18 respectively. The PSNR of the compensated image and the execution time are used as comparison criteria. As shown in figure 3.19, the PSNR obtained by using the full search method is the highest, but it requires a much longer execution time. The fast search method results shown in figure 3.20-3.23 exhibit a great reduction on computation time, with only a little loss in PSNR. The motion vector field obtained by using phase correlation in figure 3.22 clearly indicates the object movement in horizontal and vertical directions.

Figure 3.17 - The reference image used for motion estimation

Figure 3.18 - The current image

Figure 3.19 - The full search method result

Figure 3.20 - The four step search method result

Figure 3.21 - The diamond search method result

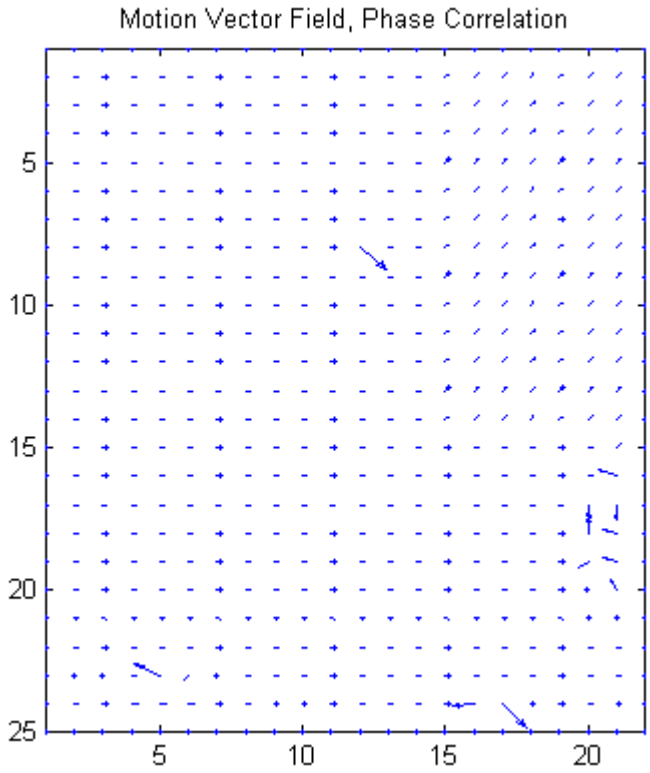


Figure 3.22 - The motion vector field by using phase correlation method

Figure 3.23 - The phase correlation method result

Caltrain	Full Search	Four Step Search	Diamond Search	Phase Correlation
PSNR (dB)	29.52	29.40	29.40	28.50
Time (Sec)	5.49	0.67	0.95	1.34

Table 3.2 - The performance of the motion estimation search method

3.5 Summary

Motion estimation is an important part of video coding; the objective of motion estimation is to find an optimal motion vector for the current macroblock in the reference frame. The reference frame is a previously coded and transmitted frame, which can be a past or future frame in display order. The best motion vector is chosen with the least cost in the search region. An accurate motion vector is able to improve the video compression quality and keep the computational complexity relatively low. The full search method is considered to be the most efficient algorithm to find the most accurate motion vector; however, the exhaustive search procedure increases the burden of computational complexity. Therefore, a lot of fast search techniques have been developed to optimize the motion estimation stage. In order to evaluate the performance of the different motion estimation search algorithms, the full search and several fast search motion estimation algorithms have been studied in this chapter. Experimental results show that the fast search algorithms are very computationally efficient compared to the exhaustive full search method, whilst increasing only a small loss in image quality.

CHAPTER 4

A SIMPLIFIED RATE CONTROL ALGORITHM FOR H.264/SVC

4.1 Introduction

Currently digital video is utilized in a very broad and increasing range of applications. The development of video compression techniques have facilitated the development of new applications including digital television, digital versatile disk (DVD) players, streaming Internet video and video conferencing. Those devices have different capabilities and transmission systems, consequently the requirements of each, with respect to spatial resolution, temporal resolution and quality are also different.

The objective of Scalable Video Coding (SVC) is the encoding of a single high-quality video bitstream, adaptable to the devices of differing capabilities. From the parent bitstream subsets may be derived which represent the sequence at a smaller frame size, lower frame rate, lesser quality, or some combination of the above. The subsets are decoded in the same manner as the parent stream, for a given instance the transmission bandwidth is reduced to that required for the corresponding subset bitstream only.

A video sequence can be encoded in several layers to meet the different requirements in frame rate, picture quality. In each case, the base layer provides basic performance, that is, lower frame rate, frame resolution or quality. The increased performance is obtained by decoding the base layer together with the enhancement layer. In a similar way, three optimal scalability modes are supported in the prior video coding standards MPEG-1 [9], H.262 MPEG-2 Video [10], H.263 [11], and MPEG-4 Visual [35]. Temporal scalability increases frame rate, spatial scalability increase frame resolution and SNR scalability increase picture quality [36]. Scalable video coding is an efficient solution to meet the different characteristics of modern video transmission systems. The variety of devices with different capabilities and connection quality can use the same high-quality video bitstream by decoding the subset bitstream in different layer.

Rate control has played an important role in video coding. The rate control in the scalable video coding is to regulate the bitstream according to the available bandwidth so that the video quality in the base and enhancement layer is kept as high as possible.

In order to meet the target bitrate and prevent the buffer from overflowing or underflowing a proper rate control scheme is important. A study of rate control algorithm in the scalable video coding is presented in the following section.

This chapter is organised as follows: The scalable video coding with three mode scalability is presented in section 4.2. The rate control scheme that used in scalable video coding is presented in section 4.3, and the related rate control method is discussed in section 4.4. The proposed algorithm is presented in section 4.5. Section 4.6 presents the experimental results and discussions. Finally, the conclusion is given in section 4.7.

4.2 H.264/ Scalable Video Coding

The objective of scalable video coding is to enable the generation of a unique bitstream that can adapt to various bitrate, transmission channels, and display capabilities and without a significant loss in coding efficiency. A success of scalable video coding standard should meet the following requirements [37]:

4.2.1 The Requirements for Scalable Video Coding

1) Keep the same coding efficiency in bitrate, quality and complexity compared to the single-layer coding.

2) Support of bitrate adaptations in each layer.

3) Support of a backward compatible in the base and enhancement layer that meet the H.264/AVC standard.

4) Support of temporal, spatial, and quality scalability.

4.2.2 The Scalability in Scalable Video Coding

The scalability is categorised in terms of temporal, spatial, and quality. The three modes of scalability are described in the following:

4.2.2.1 Temporal Scalability

A bitstream provides temporal scalability when a set of corresponding access units can be partitioned into a temporal base layer and one or more temporal enhancement layers. The subset of the bitstream in both the base and enhancement layer represents the source contents with a reduced frame rate. Temporal scalability with temporal enhancements layers as shown in figure 4.24 can provide the efficiently concept of hierarchical B-pictures [38], [39]. The T_0 is defined as the base layer and T_1, T_2 and T_3 are regarded as the enhancement layers, the enhancement layer pictures are typically coded as B-pictures.

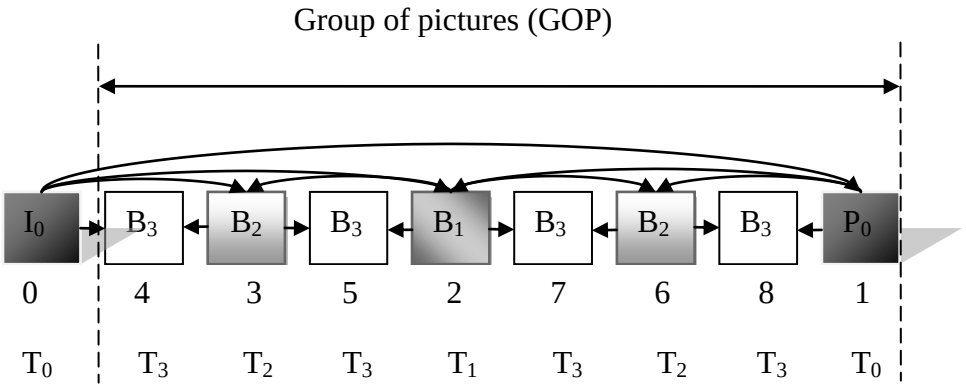


Figure 4.24 - Hierarchical B-pictures prediction structures for temporal scalability

4.2.2.2 Spatial Scalability

The bitstream in the base layer represents a low resolution of the coded frame. A higher-resolution output can be obtained by decoding the base layer combined with enhancement layers. The following steps described the procedure to encode a video sequence into two spatial layers in the MPEG-4 Visual [35] [40]:

1) Subsample the input video frame horizontally and vertically to required resolution.

2) Encode the subsample frame to form the base layer.

3) Decode the base layer fame and up-sample it to the original resolution to form a prediction frame.

4) Subtract the original frame from the prediction frame and form the residual.

5) Encode the residual to form the enhancement layer.

The spatial coding procedure is depicted in figure 4.25.

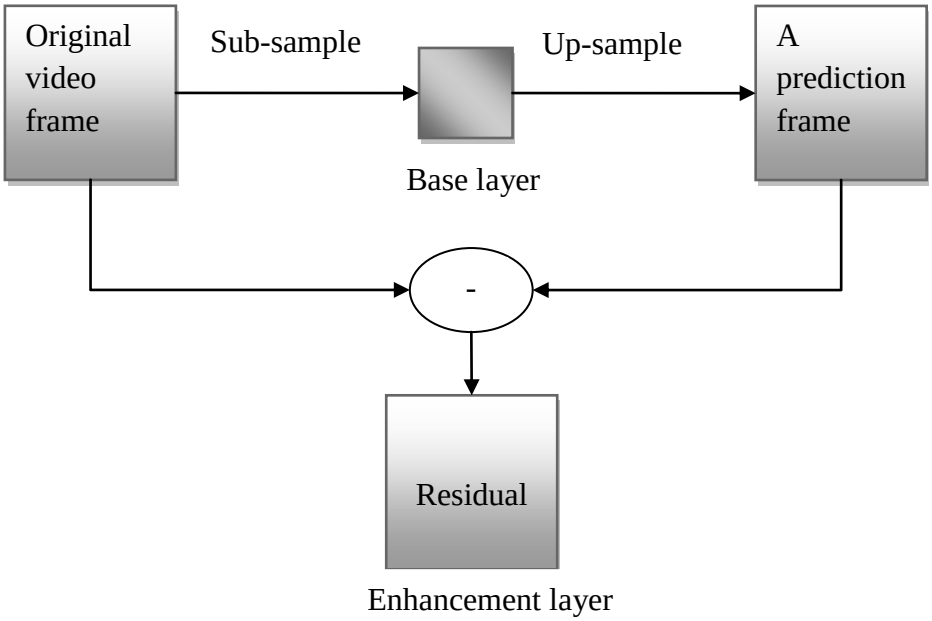


Figure 4.25 - Scalable spatial coding procedure

A lower resolution sequence can be obtained by decoding the base layer. A full-resolution sequence can be reconstructed by decoding the base layer with the enhancement layer as follows:

1) Decode the base layer frame and up-sample it to the original resolution.

2) Decode the enhancement layer residual.

3) Add the decoded enhancement residual to the decoded base layer frame to form the output frame.

Although the spatial scalability mode is supported in the previous standard [10] [11] [35], it has rarely been used, since the spatial scalability lead to a significant loss in coding efficiency and increase the complexity significantly.

The Joint Video Team of the ISO/IEC Moving Picture Experts Group and the ITU-T Video Coding Experts Group has standardized a scalable video coding extension of the H.264/AVC standard [37]. In order to improve the coding efficiency for spatial scalable coding, the additional inter-layer prediction tools have been added in scalable video coding. In each spatial layer, motion-compensated prediction and intra prediction are still employed; in addition, the inter-view prediction is used between the spatial layers for improving the rate distortion efficiency of the enhancement layers. The inter-view prediction structure is illustrated in figure 4.26.

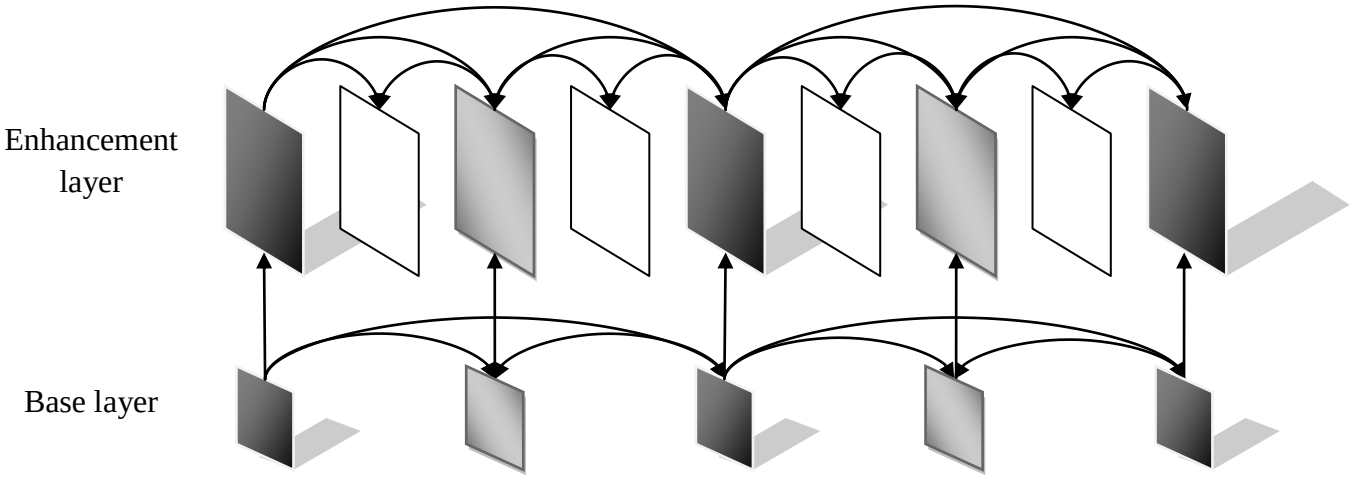


Figure 4.26 [37] - Multilayer structure with additional inter-layer prediction for enabling spatial scalable coding

4.2.2.3 Quality Scalability

Quality scalability is considered as a special case of spatial scalability with identical picture sizes for base layer and enhancement layer. The inter-layer prediction tools are also employed in the quality scalability. For the quality scalability, the subset bitstream in enhancement layer provides the same spatio-temporal resolution as the base layer, but a lower SNR compare to base layer.

4.3 Rate Control in Scalable Video Coding (SVC)

The rate control in the Joint Scalable Video Model (JSVM) [41] is to regulate the bitstream according to the available bandwidth and buffer size so that the video quality in the base and enhancement layer is kept as high as possible. In order to meet the target bitrate and prevent the buffer from overflowing or underflowing a proper rate control scheme is important. It is because that overflowing or underflowing of the buffer will result in the frame skipping and wastage of channel bandwidth. It is more channelling when a small buffer is used for a low-delay communication. As a consequence, the rate control scheme is developed along with the scalable video coding standard.

4.3.1 The Importance of Rate Control

In reality, the sequence is usually coded with variable bitrate since the video complexity is changing, however, the constraint decoder buffer size and network bandwidth required us to encode sequence at a more nearly constant bitrate. In order to meet the target bitrate, the rate control scheme is required to dynamically adjust the quantisation parameter according to the source complexity. The figure 4.27 [42] illustrated the relationship between the bitrate and quantisation parameter.

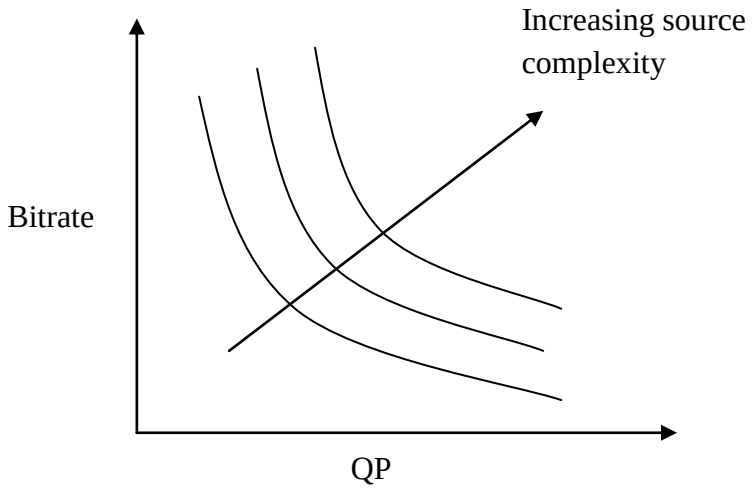


Figure 4.27 - The relationship between Bitrate and QP

The quantisation parameter can regulate how much spatial detail is retained in the decoded frame, if the quantisation parameter is set very small, then almost all the details is retained, however, a higher bitrate is required. Conversely, if the quantisation parameter is set bigger, the distortion is increased and at a loss of quality. The rate

control scheme is utilized to achieve the highest quality through adjusting the Rate-Quantization (R–Q) model when a target bitrate is given.

4.3.2 The Rate-Quantization Model

The Rate-Quantization model is used to describe the relationship between the bitrate, quantisation parameter and the video complexity. The first version of Rate-Quantization model is proposed in [43]. This model is used to calculate the corresponding quantisation parameter, and then the quantisation parameter is used for the rate distortion optimization (RDO) for each macroblock. The equation is shown as follow:

$$R = c_1 \times \frac{MAD}{QP} + c_2 \times \frac{MAD}{QP^2} + H \quad (4.1)$$

Where R indicates the total number of bits used for coding the current macroblock, MAD is computed between the current and prediction macroblock, which also indicated the complexity of the current macroblock, and H denotes the bits used for header, motion vectors and shape information. c_1 and c_2 denote the first- and the second-order coefficients, calculated based on the techniques described in [44], details are given in Appendix C. In order to simplify the training process of the Rate-Quantization model, a liner model is proposed in [45], which is shown as below:

$$R = x_1 \times \frac{MAD}{QP} + x_2 \quad (4.2)$$

Where R still indicate the total number of bits used for coding the current macroblock, but without considering the header bits. The coefficients x_1 and x_2 are updated by using the linear regression method after encoding each frame.

4.3.3 Complexity Estimation

As shown in the Rate-Quantization model, the parameter MAD is required to compute the demand quantisation parameter, however, we can only obtain the MAD after the rate distortion optimization (RDO) has used a quantisation parameter value to generate it. This problem is described as a “chicken and egg” dilemma, and then a predicted MAD

is used to solve this problem by assuming that the complexity between two pictures varies gradually. The linear prediction model [46] is proposed as below:

$$\mathrm{MAD_{cb}} = \mathrm{a_1} \times \mathrm{MAD_{pb}} + \mathrm{a_2} \quad (4.3)$$

Where $\mathrm{MAD_{cb}}$ denote MAD of the macroblock in the current frame and $\mathrm{MAD_{pb}}$denote the actual MAD in the co-located position of previous frame. The initial value of a_1 and a_2 are set to 1 and 0, and are updated after coding each macroblock. However, if the scene changed abruptly the MAD prediction by using the temporal MAD model will not accurate. A spatial MAD prediction model is proposed in [45] in addition to the temporal prediction. The $\mathrm{MAD_{rough}}$ is instead of $\mathrm{MAD_{pb}}$ to indicate the difference between the current original frame and the previous reconstructed frame. The spatial and the temporal model can be adaptively switched according to the accuracy of the MAD prediction model, which is shown as following:

$$\Gamma_{temp}[i] = \sum_{n=i-s}^{i} \left| MAD_{pred,temp}[n] - MAD_{actual}[n] \right| \quad (4.4)$$

$$\Gamma_{spat}[i] = \sum_{n=i-s}^{i} \left| MAD_{pred,spat}[n] - MAD_{actual}[n] \right| \quad (4.5)$$

Where s is the number of MAD samples that used to measure Γ. If $\Gamma_{spat}[i] > \Gamma_{temp}[i]$ then the temporal prediction will be used for the current macroblock, otherwise, the spatial prediction will be used.

4.3.4 Rate Control Scheme

With the development of the Rate-Quantization model and complexity estimation concept, the rate control can be achieved according to the following procedure as described in [46]:

1) Compute a target bit for the current frame by using the fluid traffic model and linear tracking theory [47].

2) Allocate the available bits to all non-coded basic units in the current frame equally, where the basic unit can be a frame, a slice or a macroblock.

3) Predict the MAD of current basic unit in the current frame by using the complexity estimation model.

4) Compute the quantisation parameter by using the Rate-Quantization model.

5) Perform RDO for each macroblock in the current basic unit using the quantisation parameter.

4.4 Related Rate Control Algorithm in SVC

The aim of rate control is to minimize the video distortion for a given bitrate; this can be achieved by varying the quantisation parameter according to the available bitrate and channel bandwidth. The rate control scheme can also be extended for the scalable video coding structure. A number of techniques are presented in the following to achieve this target.

In paper [47], the chicken and egg dilemma is solved by adopting a fluid-flow traffic model and a quadratic rate-distortion (R-D) model. The fluid-flow traffic model is used to determine the target bitrate for each frame according to the dynamic change of the buffer. The R-D model is used to calculate the demanded quantisation parameter to achieve the best quality. The proposed scheme is composed of two layers: group of picture (GOP) layer and frame layer, the target bitrate is first allocated to each GOP and then the bit is allocated to each frame according to the type of frame and available buffer size. In order to reduce the impact of the scene change when updating the R-D model, a sliding window mechanism is proposed. If the scene changes abruptly, a smaller window with more recently data points is used; otherwise, a window with more data points is used. The experiment results show that the average PSNR is improved and skipped frames is reduced.

A rate control scheme with buffer control and RDO consideration is proposed in [48]. The combined temporal and spatial prediction model as described in [45] is exploited to control the abrupt scene change. Besides, a rate control scheme for the hierarchical B-frames structure is proposed; the base layer is allocated with higher bitrate since it will be used as reference for motion-compensation prediction of the lower layer. The proposal of the rate control in the paper is extended to use for spatial, temporal and combined scalable enhancement layers. The experiment results show that the target bitrate can be achieved in each layer and the buffer is prevented from overflowing or underflowing.

A rate control scheme for the temporal scalability of H.264/SVC is presented in [49]. In this paper, a rate distortion model is developed based on the Cauchy-Density-Based model [50]. This model can reveal the relationship between the frame total bits, the distortion and quantisation step with more accuracy. In addition, the dependencies between temporal layers are also taken into consideration in the bit allocation step, and then more bits are allocated to the base layer. The rate control scheme can provide a good performance compared to the rate control in JVT-G012 [46] standard.

Although the proposed R-D model can solve the dilemma problem successfully in the rate control mechanism, the computational complexity is increased a lot when updating the parameters in this model. In order to decrease the coding complexity, it is necessary to simplify the model and keep the same accuracy at any time. A simplified rate control algorithm is proposed in this work [51], and the complexity is reduced successfully compared to the standard.

4.5 Proposed Algorithm

In the proposed method, the quantisation parameter can be obtained directly according to the predicted MAD. The algorithm works as following: firstly, scene changes are predicted by using the current predicted MAD and the previous actual MAD, if the predicted MAD is bigger than the previous actual MAD, the scene is regarded as change abruptly; secondly, if the scene changes abruptly then the quantisation parameter is increased directly by a threshold, otherwise, the Rate-Quantization model will be used to obtain the proper quantisation parameter for the current frame or macroblock.

In the video codec, the interdependencies of video frames are exploited for efficiency coding since the scene changed gradually in most situations. However, the scenes may change abruptly for some special case, such as a new object is occurring in the current scene. In this case, in order to suppress the visual quality “beating” or “pulsing” [52], the quantisation parameter cannot be increased too fast, the quantisation parameter is increased by a threshold in this experiment, and the max quantisation parameter change is set to 2. In order to further investigate the correlation between the quantisation parameter, the predicted MAD and the actual MAD, two experiments based on the complex sequences with high spatial detail and high amount of motion are completed,

and the experiment results are shown in figure 4.28 and figure 4.29. The figures show that the quantisation parameter changes gradually with the MAD.

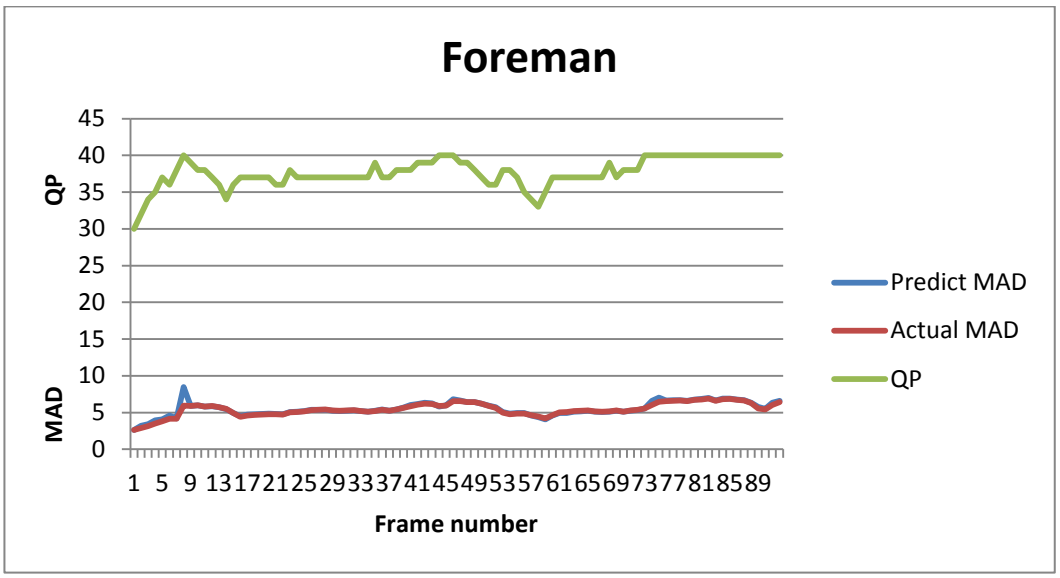


Figure 4.28 - The relationship between predict MAD, actual MAD and QP (Foreman)

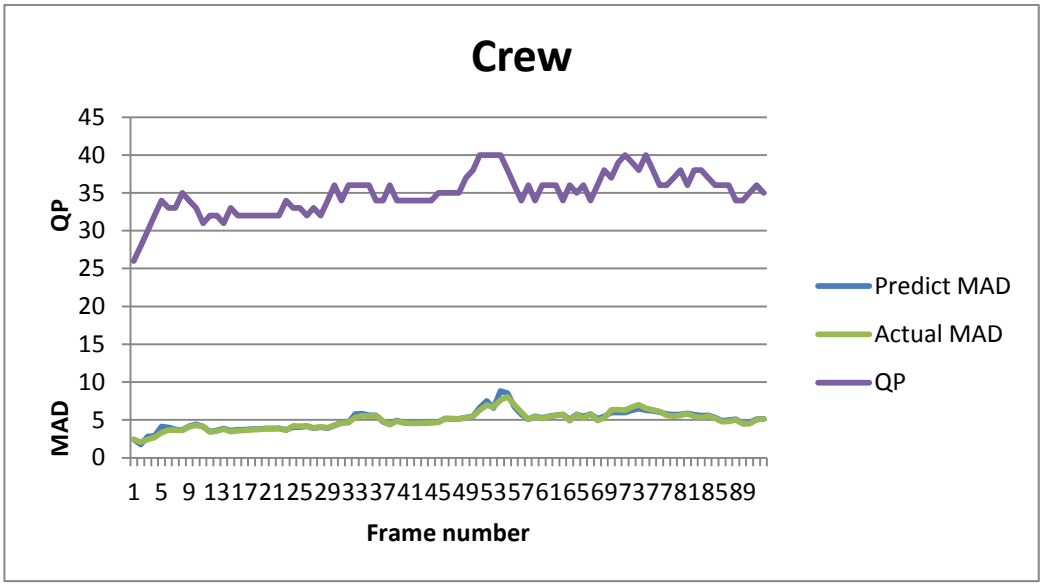


Figure 4.29 - The relationship between predict MAD, actual MAD and QP (Crew)

The block diagram of the proposed rate control scheme at the frame level is depicted in figure 4.30.

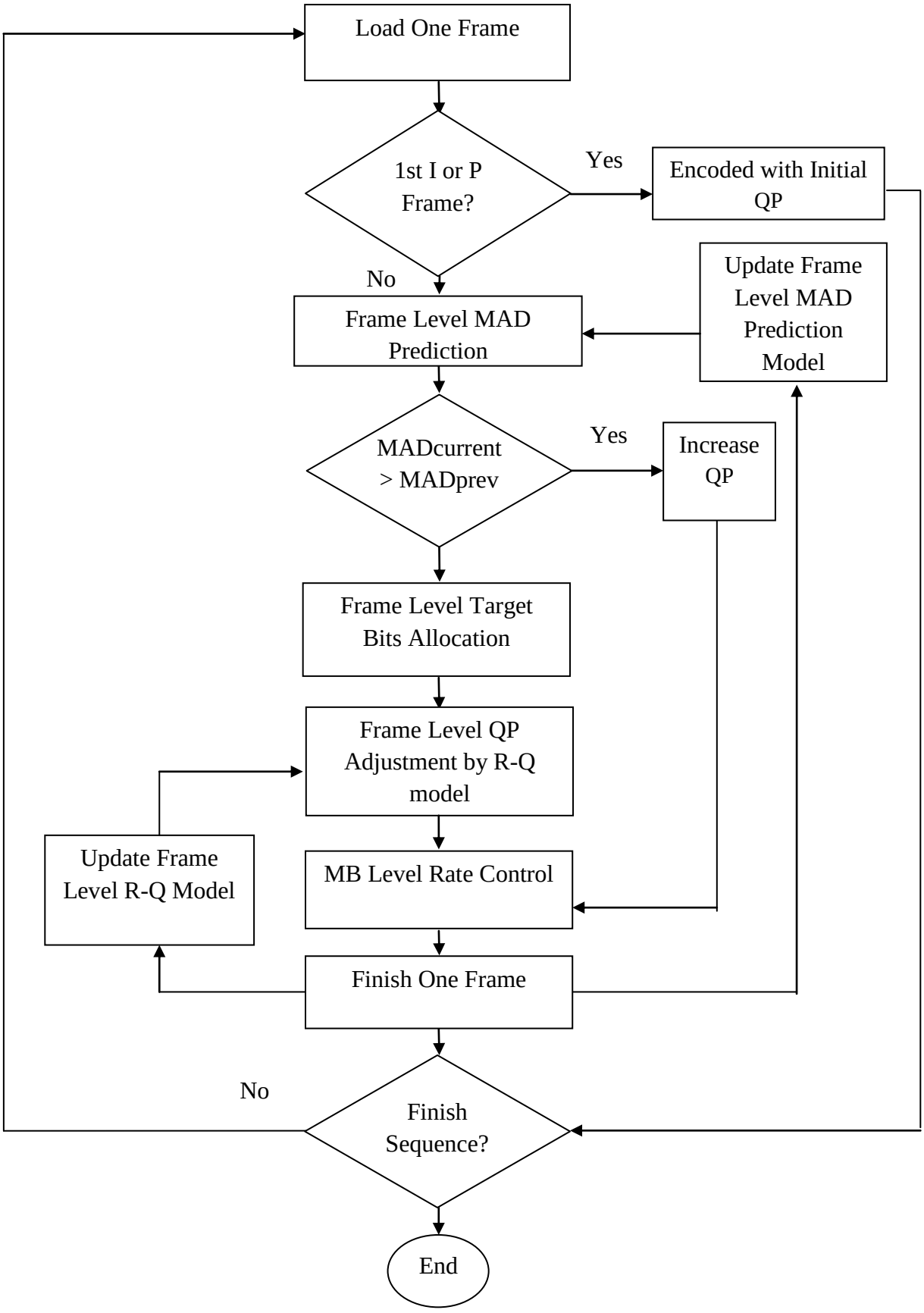


Figure 4.30 - Block diagram of the proposed rate control scheme at frame level

There are five stages for frame level rate control: 1) frame level MAD prediction. 2) frame level target bit allocation. 3) frame level quantisation parameter calculation. 4) MB level rate control if desired. 5) update the frame level Rate-Quantization model, and MAD prediction model. The description of each stage is based on the JVT-012 document [46], with the same choice of the design parameter values as the standard.

1) Frame level MAD prediction

A linear MAD prediction model is proposed to predict the MADs of current basic unit in the current frame. The linear model is proposed to solve the chicken and egg dilemma. They are updated after coding each basic unit. The details are described in section 4.3.3.

2) Frame level target bit allocation

With the leaky bucket model and linear tracking theory [47], the target bits allocated for the current frame is computed as follows:

$$f(n_{i,j}) = (1-\beta) \times \tilde{f}(n_{i,j}) + \beta \times \hat{f}(n_{i,j}) \quad (4.6)$$

where $f(n_{i,j})$ is the target sum bits allocated for the j^{th} frame in the i^{th} GOP. β is a constant and is set to 0.5 when there is no B frame and is 0.9 otherwise. β is used to specify the ratio of coding complexity between the target bits and the number of remaining bits in the current group of picture. Due to the target bit for the different frame type is varying, then the β is set to low when B frame exists.

$$\tilde{f}(n_{i,j}) = \frac{u(n_{i,j})}{F_r} + \gamma(Tbl(n_{i,j}) - B_c(n_{i,j})) \quad (4.7)$$

where $\tilde{f}(n_{i,j})$ is determined based on the target buffer level $Tbl(n_{i,j})$, actual buffer occupancy $B_c(n_{i,j})$, available channel bandwidth $u(n_{i,j})$ and frame rate F_r. γ is a constant and is set to 0.75 when there is no B frame and 0.25 otherwise. A tight buffer regulation can be achieved by choosing a large γ.

$$\hat{f}(n_{i,j}) = \frac{W_p(n_{i,j-1})T_r(n_{i,j})}{W_p(n_{i,j-1})N_{p,r}(j-1)+W_b(n_{i,j-1})N_{b,r}(j-1)} \quad (4.8)$$

where $\hat{f}(n_{i,j})$ is determined according to the remaining bits for the current frame $T_r(n_{i,j})$ in the current group of picture, the average complexity weight of P pictures and

B pictures $W_p(n_{i,j-1})$, $W_b(n_{i,j-1})$, and the number of remaining P frames and B frames $N_{p,r}(j-1), N_{b,r}(j-1)$ in each group of picture before encoding the j^{th} frame.

3) Compute the quantisation parameter and perform RDO

The quantisation parameter of the current frame is then computed by using the predicted MAD and the Rate-Quantization model. To maintain the smoothness of visual quality among frames, the quantisation parameter is adjusted by

$$Q_p[i] = \min\{Q_p[i-1] + 2, \max\{Q_p[i-1] - 2, Q_p[i]\}\} \quad (4.9)$$

where $Q_p[i-1]$ is the quantisation parameter of the previous frame. The final quantisation parameter is further bounder by

$$Q_p[i] = \min\{51, \max\{Q_p, 1\}\} \quad (4.10)$$

The quantisation parameter is then used to perform RDO, and the coding mode is selected by minimizing the following performance index:

$$J = D(s,c,MODE)|QP) + \lambda_{mode} R(s,c,MODE)|QP) \quad (4.11)$$

If the picture type is P or B and the SSD (sum of the squared difference) is used to calculate the distortion $D(s,c,MODE)|QP)$ between the original frame and its reconstruction. The lambda λ_{mode} is chosen depended on the picture type. The $R(s,c,MODE)|QP)$ denotes the total number of bits to encode the motion, header and texture information.

In the proposed method, the quantisation parameter is computed by using the MAD prediction model only. If $MAD_{current} - MAD_{previous} > 0$, the quantisation parameter is increased by a threshold directly, that is,

$$Q_p[i] = Q_p[i-1] + threshold \quad (4.12)$$

Otherwise, the quantisation parameter can be adjusted by using the Rate-Quantization model. The threshold is set to 1 in the following experiments. If the channel bandwidth is low, the threshold can be set to 2 to reduce the number of bits.

4) MB level rate control

The proposed method is based on the frame level, and then the MB level rate control is skipped by using the frame level quantisation parameter to encode all MBs within the frame. If the MB level rate control is chosen, the MAD predict model and R- Q model are updated based on the MB basis. The MB level rate control can achieve more accurate target bit matching and buffer control, but it decreases the coding efficiency.

5) Update the MAD prediction model and Rate-Quantization model

The parameters of the MAD prediction model and Rate-Quantization model are updated after encoding of each frame or MB. The details are described in APPENDIX .

4.6 Experimental Results

In order to evaluate the proposed algorithm, a comprehensive set of experiments have been carried out. The experiments are implemented on the Joint Scalable Video Model (JSVM) 9.19.9 encoder [41]. The test platform uses an Intel Core 2 CPU 6420 @ 3.20GHz with 3.0 GB RAM. The Intel VTune performance analyzer was used to measure the number of machine cycles differences which reflect the total encoding Time Saving. Additionally, bitrate and PSNR have been used to evaluate the proposed algorithm performance against the JSVM encoder.

Initial QP	28
MaxQPChange	2
Search Range	±32
NumLayers	1
Symbol Mode	CAVLC
GOP structure	IPP

Sequence	Rate Control	Frames to be encoded	Target Bitrate (kbit/s)	Actual Bitrate (kbit/s)
Foreman	Standard	100	64	65.91
	Ours	100	64	50.11
Bus	Standard	70	128	132.92
	Ours	70	128	85.28
Crew	Standard	100	96	96.78
	Ours	100	96	67.91
Soccer	Standard	100	48	53.50
	Ours	100	48	53.49
City	Standard	100	48	51.86
	Ours	100	48	32.88

Table 4.3 - Test conditions

Sequence	Rate Control	Actual Bitrate (kbit/s)	PSNR (dB)	PSNR Gain (dB)	Bitrate Increase (%)	Time Saving (%)
Foreman	Standard	65.91	32.19			
	Ours	50.11	30.25	-1.94	-6.02	+4.33
Bus	Standard	132.92	29.81			
	Ours	85.28	27.30	-2.51	-8.42	+1.40
Crew	Standard	96.78	32.36			
	Ours	67.91	30.64	-1.72	-5.31	+4.42
Soccer	Standard	53.50	29.91			
	Ours	53.49	29.85	-0.06	-2.00	+1.92
City	Standard	51.86	31.19			
	Ours	32.88	28.61	-2.58	-8.27	+4.19
Average				-1.76	-5.64	+3.25

Table 4.4 - Comparison between the proposed algorithm and the JSVM 9.19.9 software

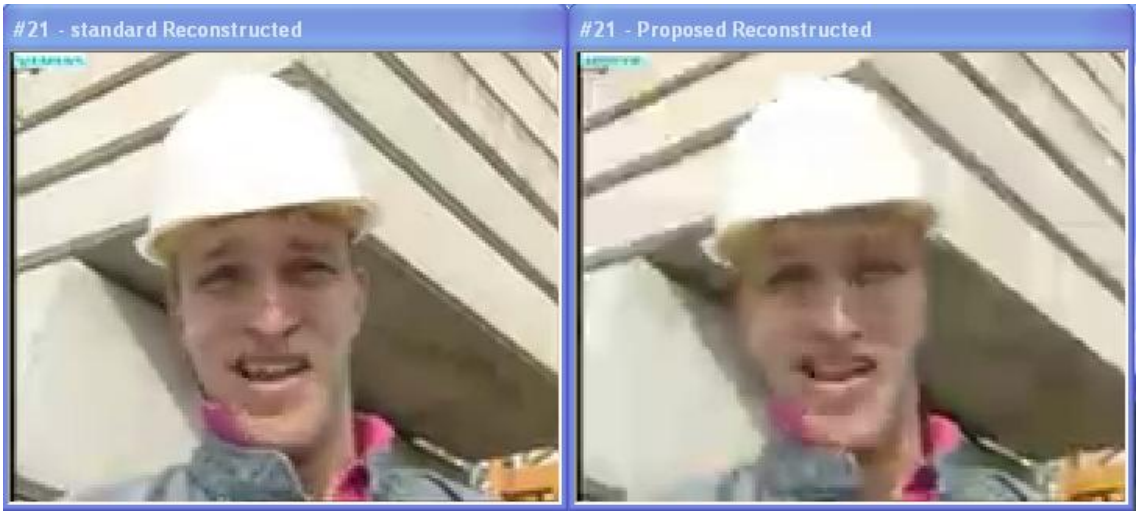


Figure 4.31 - Comparison between the standard and proposed reconstructed frame under same condition

Five standard video sequences are used in the experiments as shown in the table 4.3. From the table it can be seen that the proposed algorithm achieves an average of 3.25% time saving. However, the proposed algorithm cannot generate accurate target bitrate estimation, resulting in an average of 1.76dB losses in PSNR. In order to evaluate the subjective video quality, two frames that extracted from the standard and proposed reconstructed video sequences are shown in figure 4.31, it can be seen a quality loss in the proposed method.

A comparison between the proposed algorithm and standard with the same bitrate is shown in table 4.5. In order to make full use of the target bitrate, the initial QP in the proposed algorithm is set higher than the standard. It can be seen from the result that the average luminance PSNR of the reconstructed video is increased by up to 0.65dB at cost higher bitrate. The subjective video quality evaluation is shown in figure 4.32. It can be seen from the figure that the quality loss is negligible. The proposed scheme still achieves an average of 3.05% time saving.

MaxQPChange	2
Search Range	±32
NumLayers	1
Symbol Mode	CAVLC
GOP structure	IPP

Sequence	Rate Control	Frames to be encoded	Initial QP	Target Bitrate (kbit/s)	Actual Bitrate (kbit/s)
Foreman	Standard	100	28	64	65.91
	Ours	100	20	64	67.27
Bus	Standard	70	28	128	132.92
	Ours	70	22	128	132.6
Crew	Standard	100	28	96	96.78
	Ours	100	14	96	95.28
Soccer	Standard	100	28	48	53.5
	Ours	100	28	48	53.49
City	Standard	100	28	48	51.86
	Ours	100	28	48	49.65

Table 4.5 - Test conditions

Sequence	Rate Control	Actual Bitrate (kbit/s)	PSNR (dB)	PSNR Gain (dB)	Bitrate Increase (%)	Time Saving (%)
Foreman	Standard	65.91	32.19	-0.8	2.06	+5.44
	Ours	67.27	31.39			
Bus	Standard	132.92	29.81	-0.99	-0.24	+1.41
	Ours	132.6	28.82			
Crew	Standard	96.78	32.36	-1.15	-1.54	+3.8
	Ours	95.28	31.21			
Soccer	Standard	53.5	29.91	-0.06	-0.01	+0.77
	Ours	53.49	29.85			
City	Standard	51.86	31.19	-2.58	-4.26	+3.86
	Ours	49.65	28.61			
Average				-1.11	-0.80	+3.05

Table 4.6 - Comparison between the proposed algorithm and the JSVM 9.19.9 software with the same bitrate

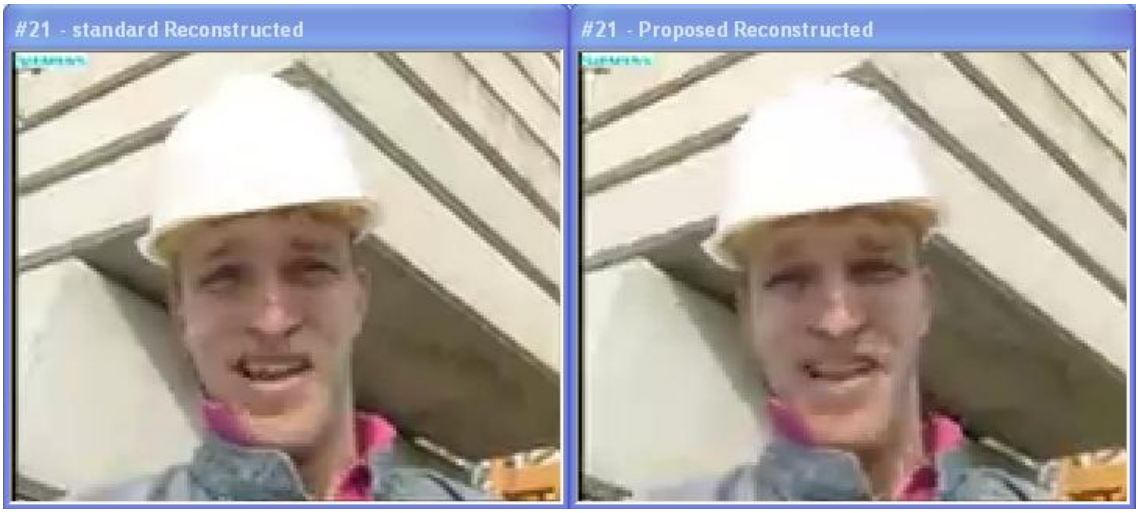


Figure 4.32 - Comparison between the standard and proposed reconstructed frame under same target bitrate

4.7 Summary

The scalable video coding technique and rate control scheme are presented in this chapter. The scalable video coding is aim to encode one single bit-stream once that meet the different requirements in frame rate, picture resolution and video quality. Three types of temporal, spatial and quality scalability are proposed in the scalable video coding to meet this requirement. The three mode scalability is discussed in details in the section 4.2.2. The rate control in the scalable video coding is aim to achieve the highest video quality by a given target bitrate in each layer. The challenge of the rate control is to prevent the decoder buffer from overflow or underflow. Since the buffer overflow will result in the frame skipping and then lost the frame. The buffer underflow will result in the bandwidth waste and the video will be delay. The rate control scheme is completed by defining a linear MAD prediction model and a Rate-Quantization model. The two models can allocate the available bitrate to each frame according to the buffer status and keep the video quality as high as possible. However, the computational complexity is increased when updating the parameters in these two models. In order to solve this problem, a simplified model is proposed in this work.

In the proposed method, the major problem of computational complexity in the scalable video coding is addressed. We proposed a direct way to obtain the quantisation parameter according to the MAD prediction scheme. The experiment results show that the proposed approach always outperforms the standard with lower bitrate and acceptable PSNR reduction and the time saving is up to 3.25%.

CHAPTER 5

INTER-VIEW REFERENCE FRAME SELECTION IN H.264/MVC

5.1 Introduction

The demand of multi-view video is increasing in the area of stereo video, free viewpoint television and multi-view television. A common element of these multimedia systems is the use of multiple views of the same scene from different viewpoints. The encoder is required to encode the N temporally synchronized video streams at the same time. Since multi-view video contains a large amount of data, a new multimedia technology is emerging to meet the high processing capability.

Multiview Video Coding (MVC) is an extension of the H.264/AVC standard that provides efficient coding of multi-view video [53]. Since all cameras capture the same scene from different viewpoints at the same time, statistical dependencies exist between adjacent views, referred to as inter-view dependencies in MVC. Combined temporal and inter-view prediction is the key for efficient MVC, where pictures cannot only be predicted from temporal reference frames, but also from corresponding inter-view reference frames in the neighbouring cameras. However, this prediction from the additional inter-view reference increases the overall encoding complexity. To further improve the coding efficiency, a number of efficient MVC techniques have been proposed which speed up the prediction process by utilizing the inter-view correlation. However, extra complex search strategies or mode decisions are needed in these methods. In this chapter, the correlation between the inter-view reference frame and the current frame is taken into consideration, and the inefficient inter-view reference frame is removed from the prediction list directly according to the phase correlation coefficients.

This chapter is organised as follows: The MVC applications and their basic structure are presented in section 5.2. The related efficient algorithm that used in MVC is summarised in section 5.3. The proposed algorithm is presented in section 5.4. Section 5.5 presents the experimental results and discussions. Finally, the summary is given in section 5.6.

5.2 H.264/Multiview Video Coding

The Joint Video Team (JVT) of the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG) has completed the standardization of MVC as an extension of H.264/AVC [54]. Multiview technology can provide a better viewing experience and additional functionality, and applications of this technology are described in the following section:

5.2.1The MVC Applications:

1) Free Viewpoint Television (FTV)

The number of views and direction can be interactively changed. The viewer can choose to watch just a single view or several views together, which may include the whole scenery.

2) Three-dimensional TV (3DTV)

3DTV can be regarded as an extension of stereoscopic view, and stereoscopic display, multi-view display or 2D-plus-depth techniques are employed in 3DTV [55], giving viewers the perception of depth.

Since the number of views is increased, video coding requires a higher compression efficiency compared to a single view. Besides compression efficiency, low delay, error resilience, and video quality should also be taken into consideration.

5.2.2 The Requirements for MVC

1) Overall video compression efficiency gain compared to simulcast single video coding

Due to the limited bandwidth condition, it is necessary to reduce the bitrate without losing video quality or increasing coding complexity, especially for low delay requirement applications.

2) Random access of a single view or frame

Due to the inter-view prediction structure used in the standard, the view dependence is increased, and the random access requires at least one intra-coded picture.

3) Scalability in both temporal and view direction

Scalability is useful in reducing the bitrate by scaling the temporal and spatial resolution. The decoder should be able to decode part of the frame to meet the low bitrate requirement.

4) Backward compatibility with the AVC

The single view bitstream extracted from the MVC sequence can meet the H.264/AVC standard.

5) Quality consistency among views

It is necessary to keep the same quality among the views.

6) Parallel encoding

Due to the picture dependence between views, parallel encoding is required for reducing the coding delay.

7) Camera parameter (extrinsic and intrinsic) transmission

The camera parameters are required to be transmitted within the bitstream in order to support intermediate view interpolation and depth perception at the decoder.

The MVC standard aims to achieve a higher compression efficiency compared to video coding of single views individually.

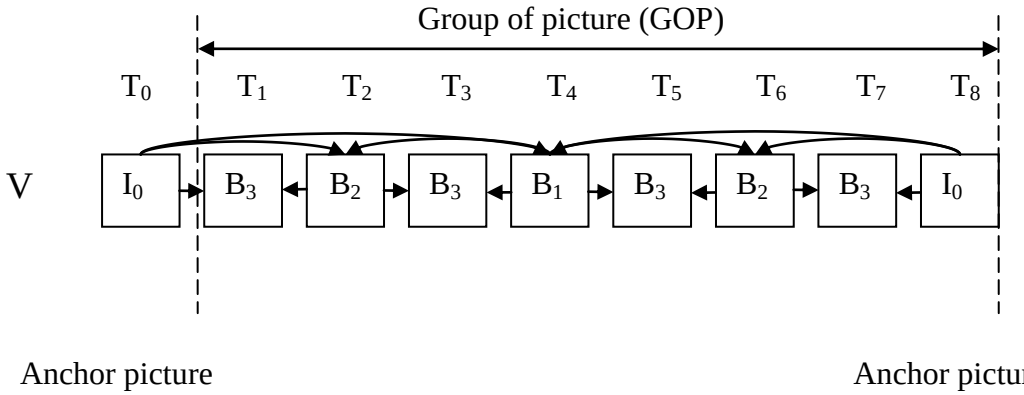
5.2.3 Prediction Structure

In the MVC standard, the hierarchical B prediction structure is used in temporal prediction, and in addition, inter-view prediction is applied to every other view, for both key and non-key pictures. The experimental results in [56] show that full inter-view prediction structure performs better than both simulcast coding and inter-view prediction for key pictures. However, due to inter-view prediction being used in

addition to temporal prediction; the prediction process is more complicated in the motion estimation stage. In order to increase the coding efficiency, the following adapted prediction structures have been implemented in the MVC.

5.2.3.1 Temporal Prediction Using Hierarchical B Pictures

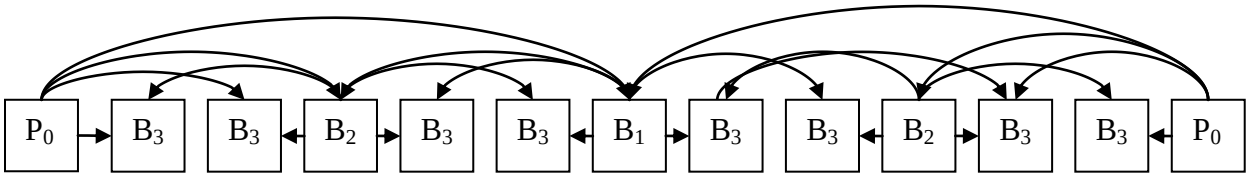
Hierarchical B pictures are used in both the AVC and MVC standards, since this prediction structure can achieve the highest coding efficiency. In the example below, a GOP (group of pictures) length of 8 is depicted in figure 5.33.



Display Order	1	2	3	4	5	6	7	8	9
Coding Order	0	5	3	6	2	7	4	8	1

Figure 5.33 - Hierarchical coding structure for temporal prediction with GOP 8

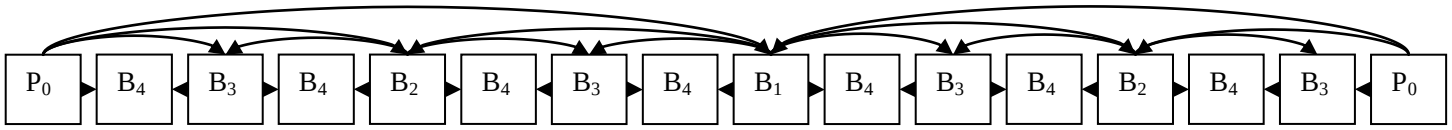
The first picture of a video sequence is always an IDR (Instantaneous Decoder Refresh) picture, which is intra-coded and consists of exactly one picture. It is called the anchor/key picture, and is coded at regular intervals, the left anchor pictures in the sequence can be either intra or inter-coded pictures [57]. A GOP is one key picture followed by a series of hierarchical B-pictures. The size of the GOP is defined in the configuration file of the MVC-encoder. The size of GOP must be equal to a power of 2, 12 or 15 for MVC, figure 5.34 and figure 5.35 depict the appropriate coding structure of GOPs with length 12 and 15.



Display Order	1	2	3	4	5	6	7	8	9	10	11	12	13
Coding Order	0	5	7	3	9	11	2	12	10	4	8	6	1

Figure 5.34 - Basic structures for coding with GOP 12

The coding efficiency can be increased by increasing the GOP size; on the other hand, the number of key pictures is reduced and the coding delay is increased.



Display Order	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Coding Order	0	9	5	11	3	13	7	15	2	14	8	12	4	10	6	1

Figure 5.35 - Basic structures for coding with GOP 15

The hierarchical B pictures prediction scheme is applied in the simulcast multi-view video coding as illustrated in figure 5.36, which shows a sequence with eight cameras and a GOP length of 8.

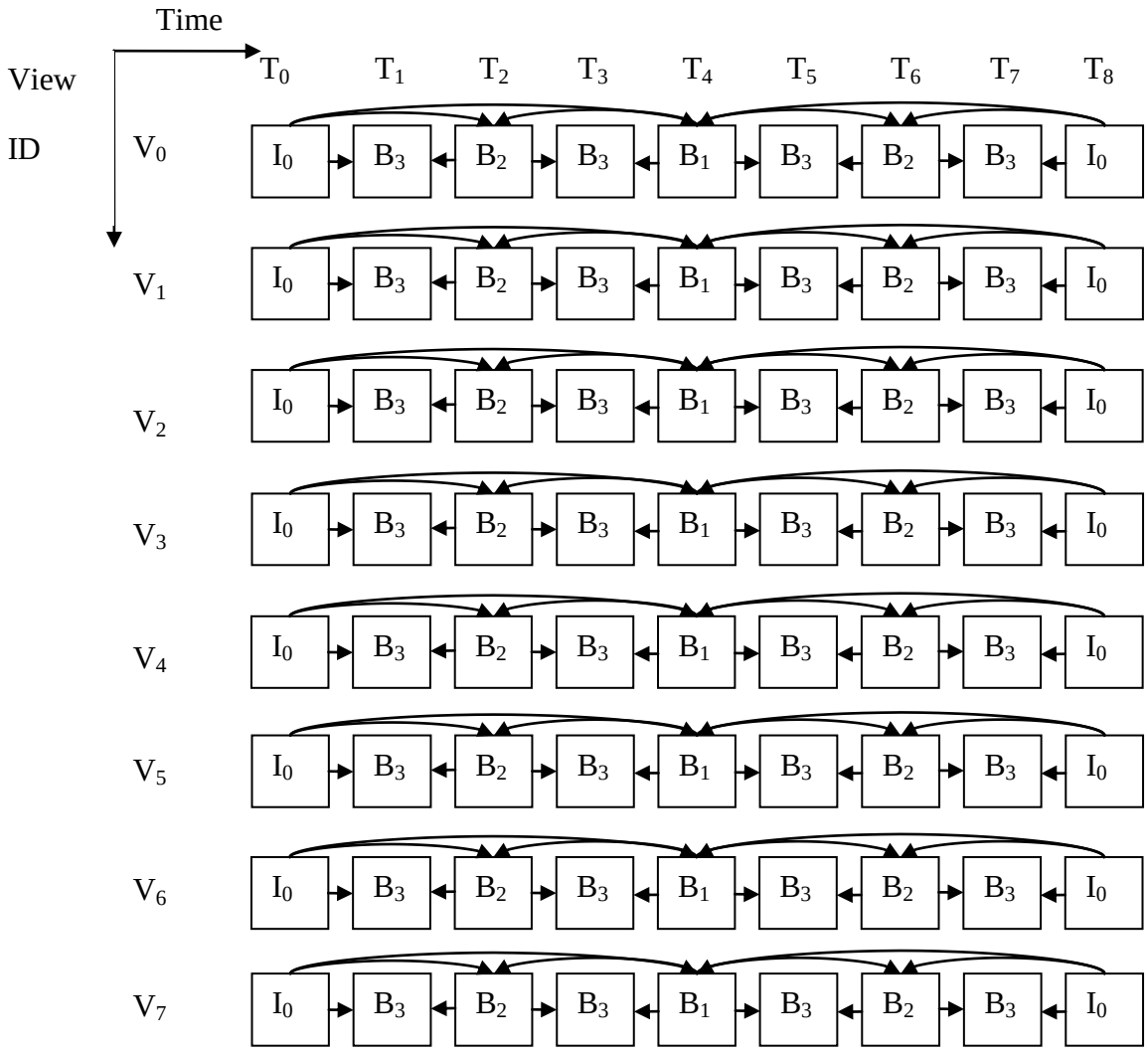


Figure 5.36 - Temporal prediction using hierarchical B pictures in the simulcast multi-view video coding

In the simulcast coding structure, each of the video sequence is encoded separately, and without correlation to other views. Although this method is simple and compatible with the H.264/AVC codec, the dependencies between views are not taken into consideration resulting in inefficiency for the MVC coding, especially on higher bitrate. In order to reduce the bitrate, the dependencies between neighbouring views are utilised in the prediction. The simulcast video coding in figure 5.36 can be adapted and extended by using inter-view prediction on the key pictures as shown in figure 5.37.

5.2.3.2 Inter-view Prediction for Key Pictures

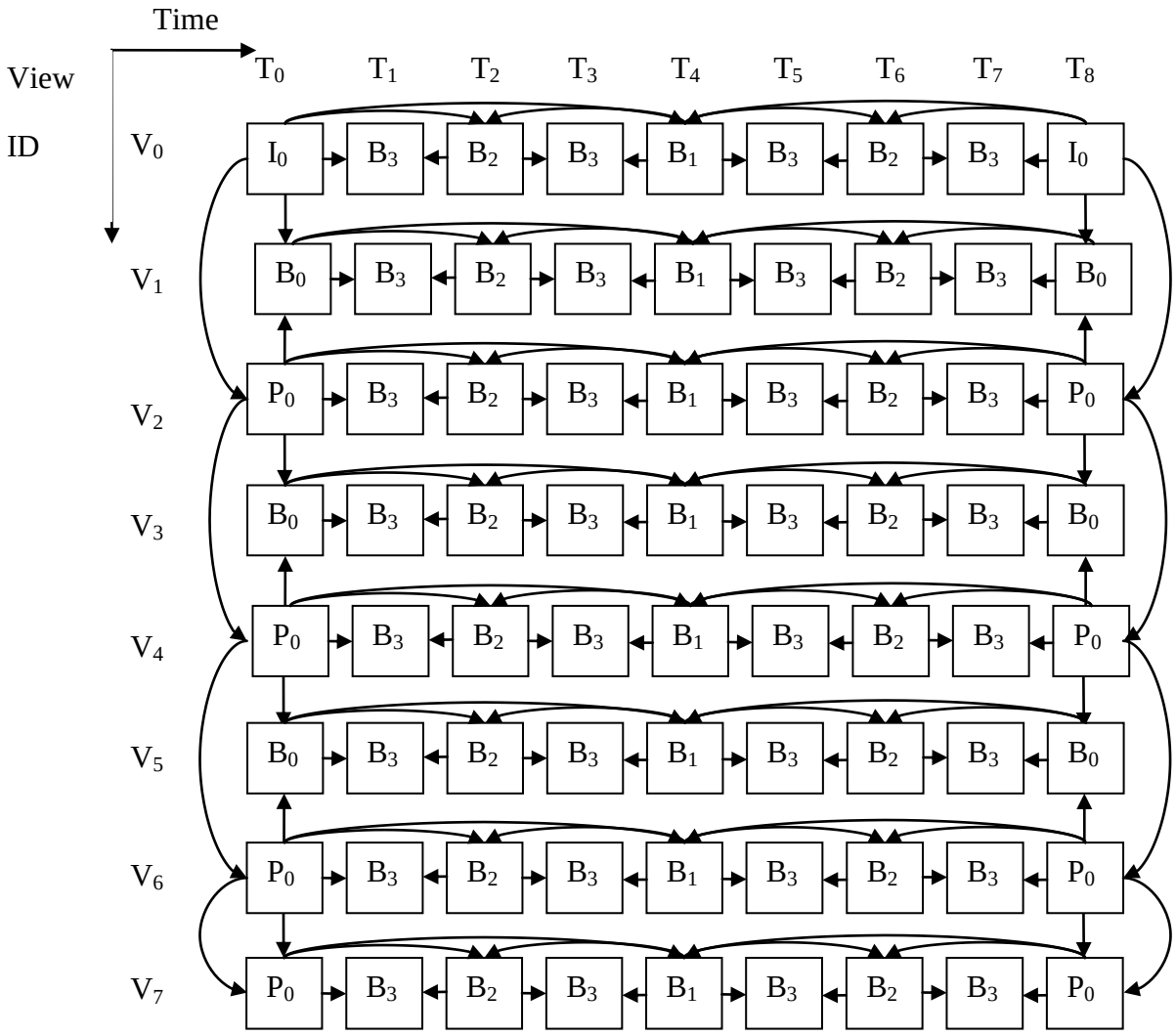


Figure 5.37 - Inter-view prediction for key pictures

V_0 is regarded as the base view. Since the first IDR pictures of the sequences except the base view are all encoded in P or B-pictures, the overall bitrate is reduced. The inter-view prediction scheme on key pictures also benefits from the multiple reference pictures in comparison to the simulcast prediction scheme. With the number of reference pictures increased, the bitrate can be further reduced in the motion estimation stage. On the other hand, the encoding complexity on the motion estimation stage is increased. In order to further improve the coding efficiency on the bitrate, an inter-view prediction scheme on both key and non-key pictures is proposed as depicted in figure 5.38.

5.2.3.3 Inter-view Prediction for Key and Non-key Pictures

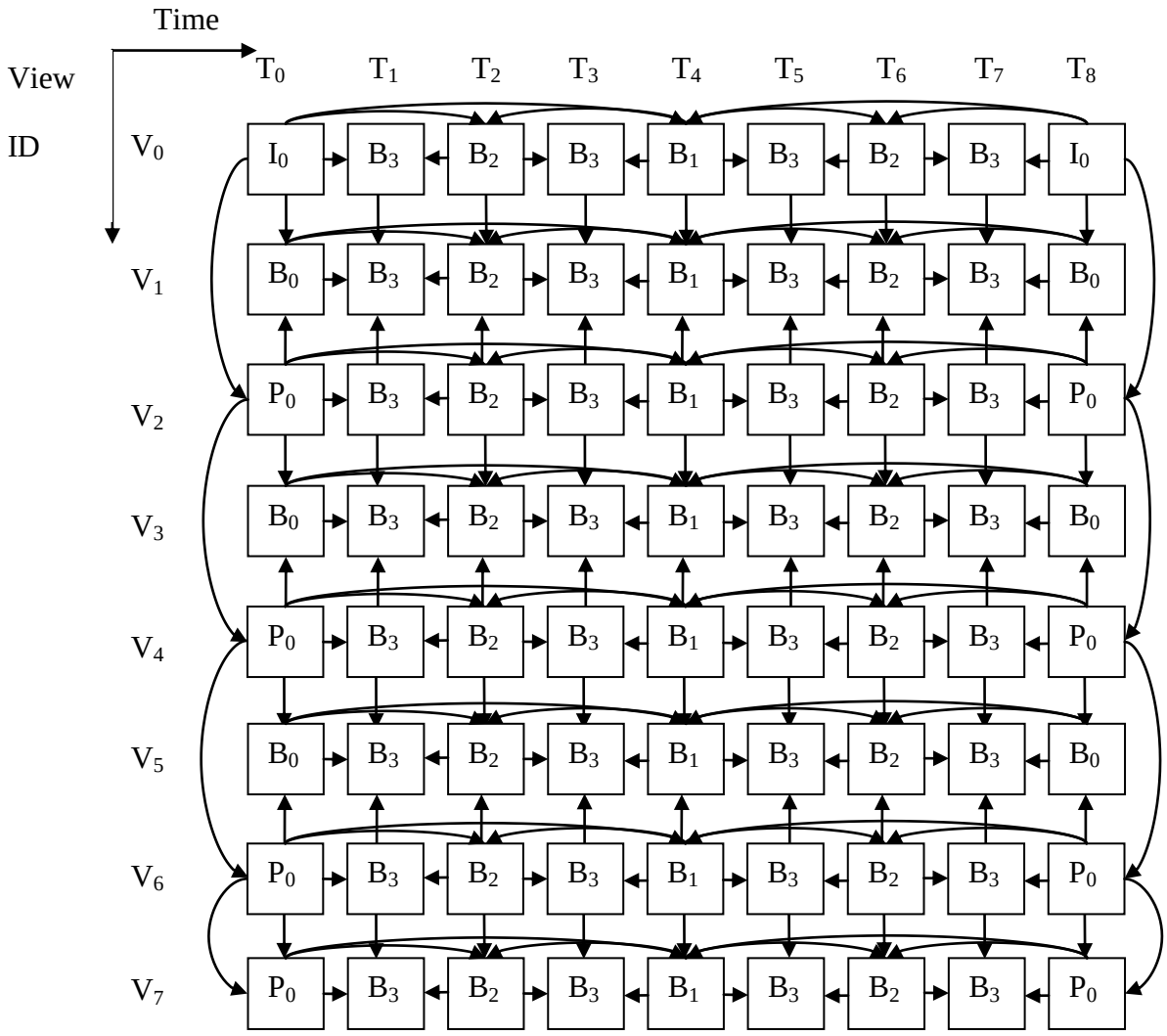


Figure 5.38 - Inter-view prediction for key and non-key pictures

The inter-view prediction for both key and non-key pictures works as depicted in figure 5.37. The only difference is that the extra inter-view pictures are added to the reference list. The non-key pictures can be predicted not only from the temporal direction, but also from the inter-view direction. The inter-view statistical dependencies of multi-view video sequences are exploited efficiently in this coding scheme. However, the large amount of reference pictures increases the coding complexity. Some related efficient algorithms have been proposed for speeding up the prediction process.

5.3 Related Efficient Algorithm for Speeding up the Prediction Process

Numerous algorithms, search methods and fast mode decision methods have been developed in MVC to speed up the prediction process.

As described in [55] [58], a fast inter-frame prediction algorithm has been developed. The prediction direction of the current macroblock is decided according to the co-located macroblocks in the temporal view. If both of the co-located macroblocks find their best match in the inter-view direction, the inter-view reference frames will be used for prediction, otherwise, the reference frames in the temporal direction will be employed. In addition, the rate distortion cost is used to estimate the performance. If the performance of the current coded macroblock is worse than a threshold, the reference frame in the other direction will be taken into consideration. In the second stage, the candidate mode is selected based on the best mode employed in the first stage. Furthermore, the motion estimation in the view direction is speeded up by utilizing the location relationship between adjacent views.

In [59], a view-adaptive mode size decision and motion search method is proposed. The mode complexity and motion homogeneity of the current macroblock are analyzed first from the previously coded macroblocks and eight neighbour macroblocks in the adjacent views, and the global disparity vector (GDV) is used to locate the corresponding macroblocks. The candidate modes are reduced based on the mode complexity analysis. The disparity search is selectively enabled, and the search range of motion estimation is dynamically adjusted according to the motion homogeneity. Experimental results show that the proposed method can significantly reduce the computational complexity of MVC.

In [60], an adaptive motion estimation and disparity estimation is proposed to reduce the computational complexity. The region homogeneity is determined by calculating the difference between the current coding macroblock and the corresponding macroblock in the reference frame. In particular, the reference frame may include temporal and inter-view references. Then, a small number of modes are selected in the rate distortion optimization process based on the region homogeneity. In addition, the search range is reduced in the inter-view and temporal-view prediction based on the location of views and region homogeneity respectively. According to [59] and [61], temporal prediction

performs better than inter-view prediction for regions with homogeneous motion, and as a consequence the disparity search for the current coding macroblock is selectively enabled according the region homogeneity behaviour.

These proposed algorithms that make use of the inter-view correlation between neighbouring views have reduced the computational complexity efficiently. However, the sophisticated calculation steps used in mode complexity and motion homogeneity decision process increase the overall computational complexity. Besides computational complexity, estimating the macroblock homogeneity by using difference and threshold has obvious problems, because the objects at different depth planes and light conditions are not taken into account in these approaches. Furthermore, the additional information of the camera geometry is required to determine the co-located position.

In the following section a phase correlation method is used to estimate the dependencies between the current frame and the inter-view reference frames. If the two frames with low correlation, then the whole inter-view reference frame will be removed from the reference list, as a result, the reference frame that used for motion estimation will be reduced. The phase correlation process only takes up 0.1% of the computational load of the whole process and do not need to care about the camera location.

5.4 Proposed Algorithm

Since the video sequences in MVC are captured by the fixed and parallel cameras, the same objects with different viewpoints are shown on adjacent views. The similarities between the current view and adjacent views are exploited for efficient prediction. Though the inter-view reference frame can be used to reduce the computational complexity, it is not always used for prediction. In order to improve the compression efficiency, an important step is to remove the less correlated inter-view reference from the reference list in the motion estimation stage.

In this section, the phase correlation method is used to determine the correlation between the current frame and the translated inter-view reference frames in the adjacent views. The phase correlation principles are described in chapter 3. In real image pairs, which contain objects at different depth planes and light conditions, perfect correlation over the entire image area will never occur [62], then sub-images (small sections of the

image) for which there will be points of strong correlation are taken into consideration in this proposed algorithm. In order to show the differences between whole frame phase correlation and sub-image phase correlation, a set of experiments have been carried out based on the sequence "Ballroom" view1. The phase correlation is calculated between the original frames in view1 and the reconstructed frames of view0 (list0), the reconstructed frames of view2 (list1). The relationship is depicted in figure 5.39.

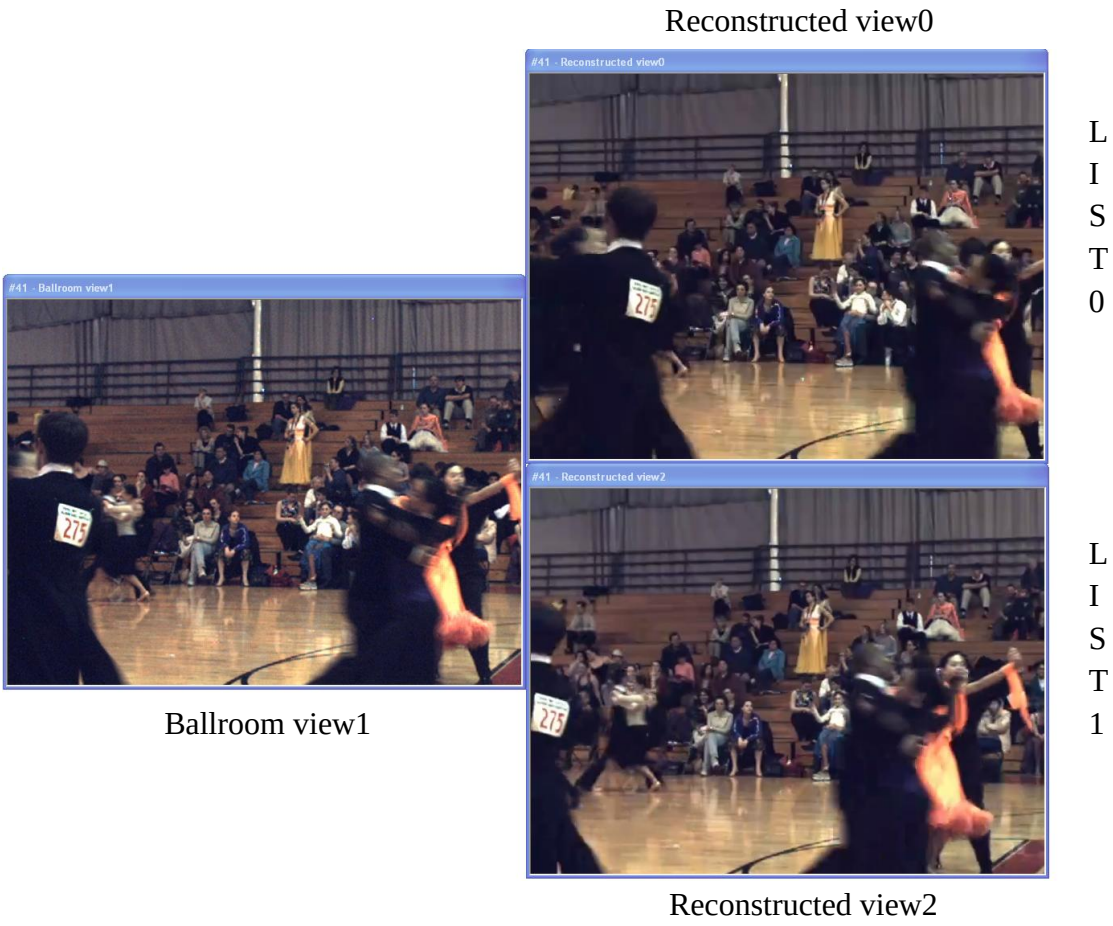


Figure 5.39 - The phase correlation between view1 and view0, view2

The experiments have been carried on 1) the whole frame 2) the whole frame (rectified), that is, the frames in view1, view0 and view2 are cropped according to the motion vectors that obtained in the whole frame testing. 3) the whole frame divided into 32×32 blocks and 4) the whole frame divided into 16×16 blocks. The test configuration is shown in table 5.6 and the test results are shown in table 5.7.

Sequence	Frames to be encoded	Image Property	Frame Rate	Basis QP
Ballroom (view1, view0 and view2)	100 Frames	640×480 (width × height)	25fps	28

Table 5.7 – The phase correlation test configuration

	Phase Correlation (Average List 0)	Phase Correlation (Average List 1)	Macroblocks (Total)
Whole Frame	0.05	0.06	100 Frames
Whole Frame (rectified)	0.16	0.15	100 Frames
MB (32×32)	0.42	0.47	30000 MBs
MB (16×16)	0.52	0.53	120000 MBs

Table 5.8 - The phase correlation test results

100 frames are encoded in each test, and the image resolution is 640×480. When the frame is divided into 32×32 blocks and 16×16 blocks, there are 300 ((640/32) × (480/32) = 300), and 1200 ((640/16) × (480/16) = 1200) blocks encoded in each frame, respectively. It can be seen from the results that the average phase correlation coefficient based on the whole frame is very small, since the sequence contains objects at different depth planes (for example the object in the green square block is not visible in view1 as shown in figure 5.40) and different light condition differences between camera sensors. In order to improve the phase correlation between the inter-view reference frames and current frames, frames are cropped according to the motion vectors obtained in the whole frame test. The cropped frames are depicted in figure 5.40, and the phase correlation was only applied to the common parts of the two frames. However, the average phase correlation with rectified field is still very low. The phase correlation with the 16×16 blocks division is the highest compared to the whole frame and 32×32 blocks. However, the small blocks cannot reflect the movement of the objects. In implementation, both the current and inter-view reference frames are divided into 32×32 blocks. The phase correlation is calculated between the current macroblock and the corresponding macroblock at the co-located position in the inter-view reference without taking into account the geometry of the cameras.

Figure 5.40 - The cropped parts of inter-view reference and current frame

The proposed inter-view reference frame skip decision strategy is shown in figure 5.41. From this figure, the inter-view reference frame skip decision is described in detail. If the phase correlation between the co-located macroblocks of current frame and inter-view reference frame is bigger than a threshold, TH_1, the current macroblock is considered to have high correlation with the macroblock in the inter-view reference frame, and then the number of Count will be increased by one, otherwise, the Count will not be changed. After comparison of all macroblocks in each frame, if Count is larger than a threshold, TH_2, the inter-view reference is considered to be highly correlated with the current frame, and then it is enabled in the prediction list, otherwise, it is disabled in the prediction list. The main difference between the proposed method and standard algorithm is that the standard algorithm always inserts the inter-view reference frames in the prediction lists, resulting in increasing computational complexity as the number of reference frame is increased. It is worthwhile to remove the inefficient inter-view reference frames from the prediction list. Based on the observation from experimental results, TH_1 is set to 0.5, which is above the average phase correlation coefficient, and TH_2 is set to 100, fixed for each quantisation parameter level and sequences with size 640×480 (300 MBs in each frame). Threshold TH_2 can be changed according to the frame size and computational complexity requirements.

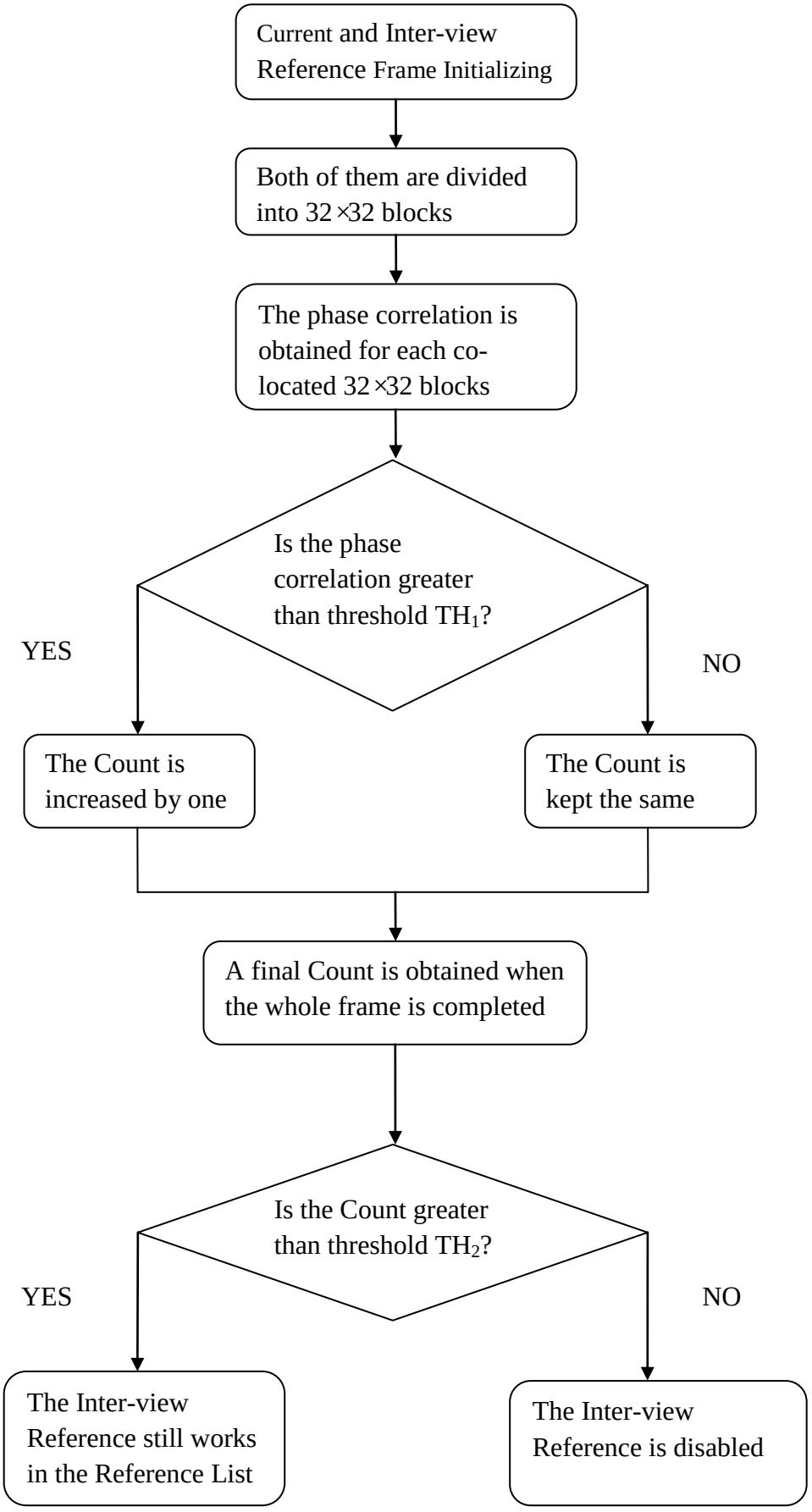


Figure 5.41 - The flow chart of the inter-view reference frame skip decision

5.5 Experimental Results

The experiments are based on the JMVC version 8.5 of MVC reference software [63]. The properties of various data sets are described in table 5.8, and all sequences are provided in the YUV 4:2:0 planar format [64]. The coding configurations settings are provided in table 5.9.

Data Set	Sequences	Image Property	Camera Arrangement
MERL	Ballroom, Exit	640x480, 25fps	8 cameras with 20cm spacing; 1D/parallel
KDDI	Race1	640x480, 30fps	8 cameras with 20cm spacing; 1D/parallel

Table 5.9 - The MVC test video sequences

	Ballroom	Exit	Race1
Frames To Be Encoded	250	250	300
No. of Viewers	8		
No. of References	3		
GOP Size	12		
QP Setting	22-27-32-37		
View Coding Order	0-2-1-4-3-6-5-7		

Table 5.10 - The configurations setting

The view coding order configuration file is summarized in table 5.10. Parameter NumViewsMinusOne specifies the number of views to be encoded minus 1, 8-view video sequence is encoded in this experiment. ViewOrder specifies the order in which the views are to be encoded. View_ID specifies the view identifier of the view to be encoded. Fwd_NumAnchorRefs specifies the number of list0 references to be used for the anchor pictures of the view identified by View_ID. It can be seen from the table that view1, view3 and view5 use two extra inter-view reference frames for prediction, which heavily increases the computational load. Views view2, view4, view6 and view7 use one more inter-view reference frame for prediction compared to the simulcast coding. The experimental results in table 5.11 – 5.13 show that the proposed method can remove the inefficient inter-view reference frame without increasing the coding efficiency. The set of BasisQP covers the range of fidelity for all test sequences.

```
#================= MULTIVIEW CODING PARAMETERS ==========================

NumViewsMinusOne                7                   # (Number of view to be coded minus 1)
ViewOrder                       0-2-1-4-3-6-5-7     # (Order in which view_ids are coded)

View_ID                         0       # view_id (0..1024): valid range
Fwd_NumAnchorRefs               0       # (number of list_0 references for anchor)
Bwd_NumAnchorRefs               0       # (number of list 1 references for anchor)
Fwd_NumNonAnchorRefs            0       # (number of list 0 references for non-anchor)
Bwd_NumNonAnchorRefs            0       # (number of list 1 references for non-anchor)

View_ID                         1       # view_id (0..1024): valid range
Fwd_NumAnchorRefs               1       # (number of list_0 references for anchor)
Bwd_NumAnchorRefs               1       # (number of list 1 references for anchor)
Fwd_NumNonAnchorRefs            1       #(number of list 0 references for non-anchor)
Bwd_NumNonAnchorRefs            1       #(number of list 1 references for non-anchor)
Fwd_AnchorRefs                  0 0     #ref_idx view_id combination
Bwd_AnchorRefs                  0 2     #ref_idx view_id combination
Fwd_NonAnchorRefs               0 0     #ref_idx view_id combination
Bwd_NonAnchorRefs               0 2     #ref_idx view_id combination

View_ID                         2
Fwd_NumAnchorRefs               1
Bwd_NumAnchorRefs               0
Fwd_NumNonAnchorRefs            0
Bwd_NumNonAnchorRefs            0
Fwd_AnchorRefs                  0 0

View_ID                         3
Fwd_NumAnchorRefs               1
Bwd_NumAnchorRefs               1
Fwd_NumNonAnchorRefs            1
Bwd_NumNonAnchorRefs            1
Fwd_AnchorRefs                  0 2
Bwd_AnchorRefs                  0 4
Fwd_NonAnchorRefs               0 2
Bwd_NonAnchorRefs               0 4

View_ID                         4
Fwd_NumAnchorRefs               1
Bwd_NumAnchorRefs               0
Fwd_NumNonAnchorRefs            0
Bwd_NumNonAnchorRefs            0
Fwd_AnchorRefs                  0 2

View_ID                         5
Fwd_NumAnchorRefs               1
Bwd_NumAnchorRefs               1
Fwd_NumNonAnchorRefs            1
Bwd_NumNonAnchorRefs            1
Fwd_AnchorRefs                  0 4
Bwd_AnchorRefs                  0 6
Fwd_NonAnchorRefs               0 4
Bwd_NonAnchorRefs               0 6

View_ID                         6
Fwd_NumAnchorRefs               1
Bwd_NumAnchorRefs               0
Fwd_NumNonAnchorRefs            0
Bwd_NumNonAnchorRefs            0
Fwd_AnchorRefs                  0 4

View_ID                         7
Fwd_NumAnchorRefs               1
Bwd_NumAnchorRefs               1
Fwd_NumNonAnchorRefs            0
Bwd_NumNonAnchorRefs            0
Fwd_AnchorRefs                  0 6
Bwd_AnchorRefs                  0
```

Table 5.11 - The view coding order relationship in MVC

The PSNR and bitrate performance over all frames in all views are reported as below:

		22		27		32		37	
Ballroom		Standard	Ours	Standard	Ours	Standard	Ours	Standard	Ours
view0	PSNR	39.265	39.2457	36.9785	36.9621	34.4536	34.4368	31.867	31.867
	Bit Rate	1605.698	1603.068	856.0896	856.104	469.2008	469.1544	267.6176	267.6176
view1	PSNR	39.04	39.0286	37.0336	37.0178	34.4957	34.4779	32.5493	32.511
	Bit Rate	1281.396	1303.938	586.9136	597.0296	288.568	292.6184	136.232	138.0816
view2	PSNR	39.3928	39.3865	37.3258	37.3174	34.7688	34.7603	32.4401	32.3749
	Bit Rate	1384.826	1423.383	688.0368	712.132	355.4136	368.388	186.396	195.1776
view3	PSNR	39.3349	39.3209	37.0171	36.9955	34.2658	34.2364	32.621	32.5403
	Bit Rate	1229.337	1283.5	563.0712	590.5176	269.9736	281.3368	129.1104	135.1
view4	PSNR	39.23	39.2264	36.9324	36.9294	34.2076	34.2097	31.9908	31.9106
	Bit Rate	1486.673	1543.398	729.5736	764.916	375.1512	394.5776	187.8712	201.8312
view5	PSNR	39.9504	39.9402	37.7254	37.7101	35.0348	35.0247	33.4207	33.3873
	Bit Rate	1122.867	1144.514	528.7392	537.5736	272.8816	276.5504	146.012	148.5816
view6	PSNR	38.7413	38.7353	36.8508	36.838	34.4852	34.4664	32.4291	32.3361
	Bit Rate	1530.086	1581.648	714.104	745.036	364.916	379.7384	193.3944	202.46
view7	PSNR	39.1816	39.1811	36.8494	36.8456	33.9873	33.9809	31.8147	31.7638
	Bit Rate	1500.338	1539.313	725.1688	749.6536	361.8768	374.592	171.9184	180.0968

Table 5.12 - The Ballroom result

		22		27		32		37	
Exit		Standard	Ours	Standard	Ours	Standard	Ours	Standard	Ours
view0	PSNR	40.1401	40.1141	38.7428	38.7282	37.0092	36.9942	34.9046	34.891
	Bit Rate	817.4448	813.1152	359.152	357.9568	191.7464	191.2088	114.9504	114.6864
view1	PSNR	39.9576	39.9515	38.5655	38.556	36.7842	36.7733	34.5702	34.5561
	Bit Rate	741.1664	757.9224	276.7448	283.048	136.2136	138.3952	77.9144	78.4816
view2	PSNR	40.1757	40.1763	38.7712	38.7686	36.9679	36.9683	34.6847	34.681
	Bit Rate	792.1128	794.072	325.6552	325.956	165.064	165.3672	95.916	96.0176
view3	PSNR	40.2229	40.2123	38.797	38.7843	36.8529	36.8356	34.4699	34.4536
	Bit Rate	694.9328	705.7832	281.4576	285.868	144.848	146.6264	85.0248	85.692
view4	PSNR	40.0057	40.0037	38.4486	38.4466	36.5347	36.5355	34.2198	34.2184
	Bit Rate	867.9936	871.7672	360.7864	361.9408	187.9512	188.2288	109.8576	110.2168
view5	PSNR	40.0726	40.0656	38.5597	38.5491	36.685	36.6705	34.4332	34.4169
	Bit Rate	807.9352	814.9472	318.5936	321.1432	161.4512	162.3696	95.78	96.1656
view6	PSNR	39.5524	39.5527	37.9707	37.9723	36.1628	36.1602	33.8852	33.8855
	Bit Rate	1071.957	1072.686	442.1048	442.9768	227.8048	228.2168	134.432	134.6088
view7	PSNR	39.991	39.9906	38.2744	38.274	36.1419	36.1417	33.6459	33.6402
	Bit Rate	945.4344	947.1112	412.6648	413.2848	209.4112	209.6088	116.4168	116.6752

Table 5.13 - The Exit result

		22		27		32		37	
Race1		Standard	Ours	Standard	Ours	Standard	Ours	Standard	Ours
view0	PSNR	40.9239	40.8878	38.1271	38.0976	35.4395	35.4234	32.9002	32.8978
	Bit Rate	1661.383	1681.55	871.8136	872.4368	469.2	468.3832	276.1848	276.256
view1	PSNR	41.3639	41.3479	38.609	38.5961	35.8275	35.8129	33.2001	33.1676
	Bit Rate	1084.611	1143.216	489.7816	519.9776	231.5536	245.7024	133.3736	143.5552
view2	PSNR	41.0088	41.0009	38.2505	38.2484	35.4755	35.4757	32.7569	32.77
	Bit Rate	1362.383	1416.1	654.0448	689.9808	315.5432	337.9416	175.7488	188.1712
view3	PSNR	41.4112	41.398	38.7867	38.7736	35.9374	35.9242	33.0375	33.0275
	Bit Rate	1062.751	1119.3	488.8488	527.7136	232.0688	255.0584	126.5048	140.484
view4	PSNR	42.0173	42.0172	39.4902	39.4835	36.9293	36.9225	34.3177	34.3138
	Bit Rate	1129.55	1137.493	599.092	607.1528	331.7368	335.8216	201.1976	204.9016
view5	PSNR	42.366	42.3623	39.8548	39.8548	37.0393	37.0538	34.0063	34.0012
	Bit Rate	889.012	914.8776	442.8704	457.0504	236.0792	246.3168	145.1136	150.0424
view6	PSNR	40.4512	40.4568	37.6711	37.6816	34.8953	34.9118	32.0413	32.0727
	Bit Rate	1659.133	1680.879	811.9672	826.4776	390.8752	401.7328	207.4632	216.6512
view7	PSNR	41.2307	41.2323	38.536	38.5369	35.7121	35.7111	32.6512	32.6583
	Bit Rate	1248.449	1251.49	575.96	580.0128	263.2776	265.8048	136.3072	137.5544

Table 5.14 - The Race1 result

The result is summarized in table 5.14. “DPSNR (dB)”, “DBR (%)”, “DT (%)” represent the average PSNR change, the percentage bitrate change, and the entire percentage coding time change respectively.

	Proposed Method Compared with Standard JMVC		
Sequences	DPSNR (dB)	DBR (%)	DT (%)
Ballroom	-0.11	3.61	21.09
Exit	-0.02	0.77	19.41
Race1	-0.15	3.82	9.44
Average	-0.09	2.73	16.65

Table 5.15 - Performance comparison between the proposed method and standard

It can be seen from table 5.14 that the proposed algorithm reduces the encoding time with a negligible loss of coding efficiency in terms of quality and bitrate. The results show that the inefficient inter-view reference frame has been removed successfully from the reference list according to the phase correlation. The proposed method has reduced the encoding time by about 16.65% on average. The average PSNR loss is 0.09 dB, and the increase of bitrate is about 2.73% on average.

5.6 Summary

In this chapter, an overview introduction to multi-view video coding is presented. Advanced MVC techniques and prediction structures are described in detail. These MVC techniques utilise the correlation relationship between the current view and adjacent views, since all the parallel cameras take the same scene from different viewpoints at the same time. With the additional inter-view references, the coding efficiency is improved significantly. On the other hand, the computational complexity is increased in the motion estimation stage. The main work in this chapter is to reduce the computational complexity, thereby making real-time applications of MVC possible.

An inter-view reference frame skip decision algorithm for MVC is presented in this chapter. The dependency relationship between the current frame and inter-view reference frame is determined based on the phase correlation of the sub-blocks. By definition, if the two frames have high correlation, then the inter-view reference frame will be enabled in the reference prediction list, otherwise, it will be disabled in the list. Experimental results show that the proposed algorithm can reduce the computational complexity of MVC without a complex search strategy or mode decision, whilst maintaining almost the same coding performance. The proposed method has been published and presented in a 2013 Data Compression Conference [65].

CHAPTER 6

CONCLUSION

6.1 Introduction

Video compression technology is widely used in the current industrialized world. Television, PCs, and video record systems rely heavily on video compression techniques. These techniques enable the data to be transmitted to homes and businesses at high-speed. At the same time, with the development of innovative techniques, storage capacity and computing power, a number of video applications has emerged, which include video conferencing, internet video streaming, high definition television and 3D television. However, the variety of video applications requires different capabilities, processing power and channel bandwidth. More advanced scalable video coding and multiview video coding techniques have been developed to meet these requirements. On the other hand, scalable and multiple video coding techniques result in increased coding complexity and a significant loss in coding efficiency. As a consequence, the future direction of video compression development is towards high efficiency video coding.

The H.264/AVC video coding standard has shown a significant improvement in video coding efficiency. The ITU-T VCEG and ISO/IEC MPEG groups have standardized a scalable video coding and multiview video coding extension of the H.264/AVC standard to meet the diverse requirements. Scalable and multiview video coding techniques are presented in this thesis. Scalability refers to a bitstream being encoded once and then adaptively decoded to meet the various needs in different devices and applications; this is achieved by removal of part of the bitstream according to the different applications' capabilities on frame rate, channel bandwidth, frame resolution and processing power. The rate control technique is utilized to keep the highest visual quality in these constrainted environments. The multiview video coding technique is used to encode multiple video sequences captured from the same scene from different viewpoints at the same time. However, the resulting high computational complexity limits its employment in real-time applications, especially on low power handheld devices. This computational complexity problem is addressed and optimized in this thesis.

This chapter summarizes the original contributions of the proposed methods. The advantages and disadvantages of the optimized methods are discussed in detail. Finally, a brief discussion of some ideas for further investigation in video coding is introduced.

6.2 Main Contributions and Results

The objective of this thesis is to reduce the computational complexity of the scalable and multiview video coding techniques of the H.264/AVC extension. The aim of the research is to keep the complexity as low as possible while keeping the bitrate low and maximising the visual quality in the constraint environment. The significant contributions of this project are mainly on a simplified rate control algorithm for scalable video coding, and complexity reduction of the multiview video coding based on phase correlation techniques. The details of the proposed algorithms and experimental results are presented in chapter 4 and chapter 5, respectively. The advantages and disadvantages of the proposed algorithms are described as follows.

6.2.1 A Simplified Rate Control Algorithm for H.264/SVC

The rate control algorithm plays an important role in the H.264/AVC scalable extension. The task of rate control is to maximize the video quality and prevent the buffer from overflowing and underflowing under the constraint of channel bandwidth, this is achieved by adjusting the quantisation parameters according to the available bits. The quantisation parameter has a great significance in rate control and coding performance. Consequently, a Rate-Quantization model is proposed to obtain the corresponding quantisation parameter. In this model, the mean absolute difference (MAD) of current frame or macroblock and available bitrate is taken into consideration. However, the MAD of the current frame or macroblock is only available after rate distortion optimization [46]. Therefore, a linear model is used to predict the MAD of current macroblock from the actual MAD of macroblock in the co-located position of previously encoded frame. On the other hand, these two models increase the computational complexity of the rate control algorithm. In order to reduce the computational complexity in the rate control stage, a simplified rate control scheme has

been proposed for temporal scalability. In video coding, in order to minimize the perceptible variations in quality after decoding, the quantisation parameter is limited to change no more than ±2 units between pictures. In addition, the quantisation parameter change is related to the MAD of the current macroblock, which has been proved in the experimental results. Based on these observations, the quantisation parameter is derived directly according to the predicted MAD of the current macroblock. That is, if the predicted MAD is changed, then the quantisation parameter is increased directly by a threshold, otherwise, the Rate-Quantization model will be used for calculating the quantisation parameter. The experimental results have demonstrated that complexity reduction is achieved by introducing this simplified rate control scheme. The advantages and disadvantages of the proposed scheme are discussed below.

Advantages

The proposed algorithm provides a simple way to control the bitrate and the objective of complexity reduction is achieved. Experimental results have shown that the proposed scheme is suitable for different type of video sequences and a time saving of up to 3.25% has been accomplished when compared to the standard.

Disadvantages

The proposed rate control scheme cannot make full use of the target bitrate, which will lead to waste of channel bandwidth. The degradation in PSNR is high when compared to the standard.

6.2.2 Inter-view Reference Frame Selection in H.264/MVC

The main problem of multi-view video coding is the huge amount of data. With the number of cameras increased, the encoder needs to encode several video sequences at the same time. In order to compress the multi-view video sequences efficiently, multiview video coding techniques have been proposed. These techniques include temporal prediction and inter-view prediction. Temporal prediction is efficient for regions with homogeneous motion, and inter-view prediction is efficient for non-homogeneous region [59] [61]. Although use of the inter-view reference picture is efficient in reducing the residual data in the motion compensation stage, the extra inter-view reference pictures result in increased computational complexity. In order to

increase the performance of multiview video coding without increasing the computational load, an inter-view reference frame selection scheme is proposed in this thesis. The dependency between the current frame and the inter-view reference frame is decided based on the phase correlation algorithm, that is, if the inter-view reference frame is highly related to the current frame, then the inter-view reference frame will still be enabled in the prediction list, otherwise, it will be removed. Since the inefficient inter-view reference frame is removed from the reference list, the number of search steps is reduced in the motion compensation stage.

Advantages

The experimental results demonstrate that the proposed algorithm based on the phase correlation algorithm is efficient on time saving. The proposed scheme could achieve a 16.65% complexity reduction without much impact on visual quality. In addition, there is no need to consider the geometry relationship between parallel cameras.

Disadvantages

Since the inter-view reference frame is disabled with the whole frame, some sub-blocks in the inter-view reference frame that are highly related to the blocks in the current frame are also disabled, which result in an increase of bitrate of about 2.73%.

6.3 Future Work

The new video coding methods such as scalable and multiview video coding have been developed to meet the current multimedia requirements. Scalability allows the single stream to be encoded once but decoded in full or part of the bitstream to meet the different requirements in video resolution, quality and frame rate. Meanwhile, a rate control scheme has been adopted in the H.264/AVC scalable extension. The objective of rate control is to regulate the bitrate to maximize the video quality under the constraint of channel bandwidth, and to prevent the buffer from overflowing and underflowing. However, these improvements come at the price of an increased complexity. In chapter 4, a simplified rate control algorithm has been proposed to reduce the computational complexity. The optimized algorithm can achieve a time saving of about 3.25% compared to the standard. However, this algorithm cannot accurately allocate the target bitrate, resulting in an average of 1.76 dB losses in video

quality. The basic unit is selected as a frame in our proposed method; however, there exists a bit allocation error between the actual bitrate and target bitrate. This problem can be overcome by using a smaller basic unit. The basic unit can also be a macroblock, a slice or a field. If the basic unit is selected as a macroblock, a slice or a field, then a more accurate bit allocation scheme can be achieved. The basic unit rate control works same as the frame layer, and the basic unit layer rate control selects the values of quantisation parameters for all basic units in a frame, so that the sum of generated bits is close to the target bits. On the other hand, it will further increase the computational complexity. For the overall rate control scheme as described in the standard [46], an important stage is to perform RDO for all MBs in the current basic unit. For this calculation of the Rate-Quantization model is needed for each basic unit, which can heavily increase computational requirements. Likewise, in the basic unit layer, in order to reduce the blocking artifacts and maintain the smoothness of visual quality between basic units, the quantisation parameter is bounded by a range. Based on our observations, the value of quantisation parameter for the current basic unit can be decided directly according to the MAD prediction model. The process is the same as that at the frame layer. Further research will be carried out to decide the number of MBs that constitute a rate control basic unit.

Multi-view video is a collection of multiple videos captured by several cameras from different viewpoints at the same time. One major goal in the future would likely be to encode multiple sequences in 3D. Multi-view videos in 3D can provide a more realistic video experience to users. However, the huge amount of data required for multi-view videos limits the development of 3D television. The computational complexity issue is emphasized in the current stage of work. In chapter 5, an inter-view reference frame selection scheme has been proposed to reduce the computational complexity successfully. However, the bitrate is increased about 2.73% on average. Further research should be carried out to reduce the bitrate. Since the whole inter-view reference frame is removed from the prediction list, some sub-blocks in the inter-view reference frame that are highly correlated to current frame are also disabled. A more efficient approach would be to retain the highly correlated sub-blocks in the prediction list. The phase correlation algorithm can be further extended to decide the search range and whether the sub-blocks should be used for prediction or not in the inter-view reference frame.

6.4 Summary

A brief history of video coding standards is described in chapter 1. The basic knowledge of video coding concepts and related advanced innovative techniques are introduced in chapter 2. Video coding techniques such as motion compensation, DCT, quantization, and entropy coding are described in detail. An innovative phase correlation and several fast search motion estimation technique algorithms are compared to the full search algorithm in chapter 3. The full search motion estimation can always find its best match with the least cost; however, it is computationally intensive and difficult to implement in real time applications. The phase correlation and several fast search algorithms have been developed to reduce the computational complexity. However, the fast search algorithms cannot guarantee to find the best match with the least cost, and they usually give poorer compression performance than full search algorithms. The comparative results are shown in table 3.2.

The scalable video coding and rate control algorithms are described in chapter 4. The rate control in the scalable video coding is aimed at generating a bitstream that meets the target bitrate and maintains the video quality as high as possible in each layer. A MAD prediction model and a Rate-Quantization model are used to solving the chicken and egg dilemma. The relationship between the predicted MAD, actual MAD and quantisation parameter are analyzed in section 4.5. A novel algorithm was developed, based on these observations, in which the quantisation parameter can be obtained directly rather than via calculation of the Rate-Quantization model. The experimental results show that the computational complexity of the rate control algorithm is reduced successfully, which enables it to be used in real time application.

The multi-view video coding applications, requirements and prediction structures are described in detail in chapter 5. Multiview video coding techniques make use of the inter-view reference frames in the adjacent view for prediction, which is efficient in reducing the bitrate and improving the visual quality. On the other hand, the additional inter-view reference frame requires a relatively high-complexity algorithm in the motion compensation stage. The phase correlation algorithm presented in chapter 2 is used to remove the inefficient inter-view reference frame from the prediction list. If the inter-view reference frame is highly related to the current frame, then it will be used for prediction, otherwise, it will be removed from the prediction list. The phase correlation

algorithm is carried on the smaller block size (32×32) rather than the whole frame. Based on the experimental results summarised in table 5.8; two thresholds are used for the decision of whether to take into account the inefficient inter-view reference frame. The experimental results in table 5.15 show that the proposed algorithm successfully reduces the computational complexity without significant loss of visual quality.

Chapter 6 summarizes the original contributions in scalable and multiview video coding. The proposed methods are efficient in reducing the computational complexity, and the optimized algorithms can be useful in low power handheld devices and real time applications.

References

[1] “International Organization for Standardization.” Internet: http://www.iso.org/. [February 2013]

[2] “International Electrotechnical Commission.” Internet: http://www.iec.ch/. [February 2013]

[3] “International Telecommunications Union.” Internet: http://www.itu.int/ITU-T/. [February 2013]

[4] http://en.wikipedia.org/wiki/ISO/IEC_JTC1. [February 2013]

[5] http://mpeg.chiariglione.org/. [February 2013]

[6] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the High Efficiency Video Coding (HEVC) standard,” IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1648–1667, Dec. 2012.

[7] http://en.wikipedia.org/wiki/H.120. [February 2013]

[8] ITU-T: Line Transmission on Non-telephone Signals: Video Codec for Audiovisual Services at p×64 kbit/s. CCITT Recommendation H.261, 1990.

[9] Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s—Part 2: Video, ISO/IEC 11172-2 (MPEG-1 Video), ISO/IEC JTC 1, Mar. 1993.

[10] Generic Coding of Moving Pictures and Associated Audio Information— Part 2: Video, ITU-T Rec. H.262 and ISO/IEC 13818-2 (MPEG-2 Video), ITU-T and ISO/IEC JTC 1, Nov. 1994.

[11] Video Coding for Low Bit Rate communication, ITU-T Rec. H.263, ITU-T, Version 1: Nov. 1995, Version 2: Jan. 1998, Version 3: Nov. 2000.

[12] MPEG-4: International Standard ISO/IEC JTC1/SC29/WG11 N4030, Coding of Moving Pictures Audio, 2001.

[13] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A.Luthra. “Overview of the H.264/AVC Video Coding Standard” IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 7, PP. 560-576, Jul. 2003.

[14] Ohm, J.; Sullivan, G.J., "High efficiency video coding: the next frontier in video compression [Standards in a Nutshell]," Signal Processing Magazine, IEEE , vol.30, no.1, pp.152,158, Jan. 2013.

[15] S. Wenger,M.M. Hannuksela,T. Stockhammer, M. Westerlund, D. Singer. RFC 3984: RTP Payload Format for H.264 Video. p. 2, February 2005.

[16] ISO/IEC 14496-10 and ITU-T Rec. H.264, Advanced Video Coding, 2003.

[17] Iain E.G. Richardson, Video Codec Design, Thomson Press (India) Ltd., New Delhi, England. P58 (2002).

[18] Chad Fogg,Didier J. Legall,Joan L. Mitchell,William B. Pennebaker. MPEG Video Compression Standard. ISBN 13: 9780412087714.October 1996, P52.

[19] V. Bhaskaran and K. Konstantinides. Image and Video Compression Standards -Algoritms and Architectures. Boston, MA: Kluwer Academic Publishers, 1997.

[20] Iain E.G. Richardson, H.264 and MPEG-4 Video Compression, Thomson Press (India) Ltd., New Delhi, England. P15 (2003).

[21] Iain E.G. Richardson, H.264 and MPEG-4 Video Compression, Thomson Press (India) Ltd., New Delhi, England. P171 (2003).

[22] K. P. Lim, G. Sullivan, and T. Wiegand “Text Description of Joint Model Reference Encoding Methods and Decoding Concealment Methods.” Document JVT-N046, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, Hong Kong, Jan. 2005.

[23] J. Ostermann, J. Bormans, P. List, D. Marpe, M. Narroschke, F. Pereira, T. Stockammer and T. Wedi. “Video Coding with H.264/AVC: Tools, Performance, and Complexity.” IEEE Circuits and Systems, vol. 4, no. 1, April 2004.

[24] Iain E.G. Richardson, H.264 and MPEG-4 Video Compression, Thomson Press (India) Ltd., New Delhi, England. P52 (2003).

[25] G. Bjontegarrd and K. Lillevold. “Context-adaptive VLC coding of coefficients.” in JVT Document, vol. JVT-C028, Fairfax, Virginia, USA, May 2002.

[26] D. Marpe, H. Schwarz, G. Bldttermann, G. Heising, and T. Wiegand, “Context-based Adaptive Binary Arithmetic Coding in JVT/H.26L" in Proceedings of International Conference on Image Processing, Rochester, NY, USA, pp. 513-516, 2001.

[27] Iain E.G. Richardson, H.264 and MPEG-4 Video Compression, Thomson Press (India) Ltd., New Delhi, England. (2003).

[28] J. R. Jain and A. K. Jain, Displacement measurement and its application in interframe image coding, IEEE Trans. Commun., 29, December 1981.

[29] M. Ghanbari, The cross-search algorithm for motion estimation, IEEE Trans. Commun., 38, July 1990.

[30] Lai-Man Po, and Wing-Chung Ma, “A Novel Four-Step Search Algorithm for Fast Block Motion Estimation”, IEEE Trans. Circuits And Systems For Video Technology, vol 6, no. 3, pp. 313-317, June 1996.

[31] S. Zhu and K.-K. Ma, “A new diamond search algorithm for fast block-matching motion estimation”, IEEE Trans. Image Process., vol. 9, no. 2, pp. 287-290, Feb. 2000.

[32] J. J. Pearson, D. C. Hines, S. Goldman, and C. D. Kuglin. “Video rate image correlation processor,” in Proc. SPIE, Application of Digital Image Processing, vol. 119, pp 197-205, August 1977.

[33] International Telecommunication Union. ITU-T Recommendation P.800.1: Mean Opinion Score (MOS) terminology. Technical report, July 2006.

[34] ITU-R Recommendation BT.500-11, “Methodology for the subjective assessment of the quality of television pictures,” Geneva, 2002.

[35] Coding of audio-visual objects—Part 2: Visual, ISO/IEC 14492-2 (MPEG-4 Visual), ISO/IEC JTC 1, Version 1: Apr. 1999, Version 2: Feb. 2000, Version 3: May 2004.

[36] Iain E.G. Richardson, Video Codec Design, Thomson Press (India) Ltd., New Delhi, England. P84 (2002).

[37] H. Schwarz, D. Marpe, and T. Wiegand. “Overview of the Scalable Video Coding Extension of the H.264/AVC Standard,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 9, pp. 1103-1120, Sep. 2007.

[38] H. Schwarz, D. Marpe, and T.Wiegand, Hierarchical B Pictures, Joint Video Team, Doc. JVT-P014, Jul. 2005.

[39] H. Schwarz, D. Marpe, and T. Wiegand, “Analysis of hierarchical B-pictures and MCTF,” in Proc. ICME, Toronto, ON, Canada, Jul. 2006, pp. 1929–1932.

[40] Iain E.G. Richardson, H.264 and MPEG-4 Video Compression, Thomson Press (India) Ltd., New Delhi, England. P142 (2003).

[41] JSVM 9.19.9 Software Package, CVS server for the JSVM software, April.2010.

[42] http://www.pixeltools.com/rate_control_paper.html. [February 2013]

[43] H.J.Lee and T.H.Chiang and Y.Q.Zhang. “Scalable Rate Control for MPEG-4 Video”. IEEE Trans. Circuit Syst. Video Technology, 10: 878-894, 2000.

[44] T. Chiang and Y.-Q. Zhang, “A new rate control scheme using quadratic rate distortion model,” IEEE Trans. Circuits Syst. Video Technol., pp. 246–250, Feb. 1997.

[45] 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, pp. 68–78, Jan. 2007.

[46] Z. G. Li, F. Pan, K. P. Lim, G. Feng, X. Lin, and S. Rahardja, “Adaptive basic unit layer rate control for JVT,” presented at the 7th JVT Meeting, Pattaya II JVT-G012-rl Thailand, Mar. 2003.

[47] Z. G. Li, Lin Xiao, C. Zhu and Pan Feng. “A Novel Rate Control Scheme for Video over the Internet”. In Proceedings ICASSP 2002, Florida, USA, 2065--2068, May 13-17, 2002.

[48] Yang Liu, ZhengGuo Li, Yeng Chai Soh, “Rate control of H.264/AVC scalable extension”, IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 1, pp. 116-121, 2008.1.

[49] ZHU Tao, ZHANG Xiong-wei, WANG Jin-ming, HUANG Jian-jun, " Rate Control Scheme for Temporal Scalability of H.264/SVC Based on a New Rate Distortion Model" Journal of Convergence Information Technology, Volume 6, Number 1. January 2011.

[50] Nejat Kamaci, Yucel Altunbasak, Russell M. Mersereau, "Frame bit allocation for the H.264 AVC video coder via Cauchy-density-based rate and distortion models", IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 8, pp. 994-1006, 2005.8.

[51] Guang Y. Zhang, Abdelrahman Abdelazim, Stephen James Mein, Martin Roy Varley and Djamel Ait-Boudaoud, "A simplified rate control algorithm for H.264/SVC", SPIE conference, April 2011.

[52] Athanasios Leontaris, Alexis Michael Tourapis, Dolby Laboratories, "Rate Control reorganization in the Joint Model (JM) reference software," Joint Video Team of ISO/IEC MPEG and ITU-T VCEG, JVT-W042, April, 2007.

[53] Aljoscha Smolic, "Introduction to Multiview Video Coding", ISO/IEC JTC 1/SC 29/WG 11N9580, Antalya, Turkey, January 2008.

[54] Information technology-Coding of audio-visual objects-Part 10: advanced video coding, Amendment 1: multiview video coding ISO/IEC 14496-10: 2008/FDAM 1:2008(E), ISO/IEC JTC1 doc ref SC29N 9783 (FDAM), March 2009.

[55] http://en.wikipedia.org/wiki/3D_television.

[56] Merkle, P., Smolic, A., Mueller, K., Wiegand, T., 2007. Efficient prediction structures for multi-view video coding. IEEE Trans. on Circuits Syst. Video Technol., 17(11):1461-1473.

[57] JMVC 8.5 (Joint Multiview Video Coding) software manual, March 26, 2011

[58] X. M. Li, D. B. Zhao, X. Y. Ji, Q. Wang, and W. Gao, "A fast inter frame prediction algorithm for multiview video coding," in Proc. IEEE Int. Conf. Image Process. (ICIP), vol. 3, Sep. 2007, pp. 417–420.

[59] L. Shen, Z. Liu, T.Yan, Z. Zhang, and P. An, "View-Adaptive Motion Estimation and Disparity Estimation for Low Complexity Multiview Video

Coding," IEEE Trans. Circuits Syst. Video Technol., vol. 20, no. 6, pp. 925-930, Jun. 2010.

[60] Pan Gao; Qiang Peng; Qionghua Wang; Xiangkai Liu; Chengde Zhang; , "Adaptive disparity and motion estimation for Multiview Video Coding," Image and Signal Processing (CISP), 2011 4th International Congress on , vol.1, no., pp.66-71, 15-17 Oct. 2011.

[61] X. M. Li, D. B. Zhao, S. W. Ma, and W. Gao, "Fast disparity and motion estimation based on correlations for multiview video coding," IEEE Trans. Consumer Electron., vol. 54, no. 4, pp. 2037-2044, Nov. 2008.

[62] M. Lukacs, "Predictive coding of multi-viewpoint image sets," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Tokyo, Apr. 1986, pp. 521–524.

[63] JMVC (Joint Multiview Video Coding) software 8.5., Mar. 26, 2011.

[64] Y. Su, A. Vetro, and A. Smolic (July, 2006). Common Test Conditions for Multiview Video Coding ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Doc. JVT-T207.

[65] Guang Y. Zhang, Abdelrahman Abdelazim, Stephen James Mein, Martin Roy Varley, and Djamel Ait-Boudaoud, "Inter-view Reference Frame Selection in Multi-view Video Coding", Data Compression Conference, Snowbird, Utah, March 20 - 22, 2013. (Accepted as a poster)

APPENDIX A

HIGH-COMPLEXITY MOTION ESTIMATION AND MODE DECISION IN H.264/AVC

A.1 Overview

In H.264, many Intra and Inter modes are used for encoding a macroblock. The best mode is chosen that gives the smallest cost in the Rate Distortion Optimization (RDO) process, and the Lagrangian cost function (Equation 2.3) is used to compute the cost for each mode. The reconstructed video quality is the best by using this exhaustive search algorithm; on the other hand, it is the most time consuming process of the encoder.

A.2 The Lagrangian Cost

The best mode for coding the macroblock is done by minimizing the Lagrangian function as follows [1]:

$$\mathrm{J}(\mathrm{s},\mathrm{c},\mathrm{MODE}|\mathrm{QP},\lambda_{\mathrm{MODE}}) = \mathrm{SSD}(\mathrm{s},\mathrm{c},\mathrm{MODE}|\mathrm{QP}) + \lambda_{\mathrm{MODE}} \cdot \mathrm{R}(\mathrm{s},\mathrm{c},\mathrm{MODE}|\mathrm{QP}) \quad \text{(A.1)}$$

where J indicates the cost for each pair of given parameters, that is, macroblock quantiser QP, Lagrange multiplier λ_{MODE} for mode decision, $\mathrm{R}(\mathrm{s},\mathrm{c},\mathrm{MODE}|\mathrm{QP})$ is the number of bits associated with choosing MODE and QP, which include the bits for the header, motion vector and reference picture, and the transformed coefficients.

The $\mathrm{SSD}(\mathrm{s},\mathrm{c},\mathrm{MODE}|\mathrm{QP})$ is the sum of the squared differences between the original block s and its reconstruction c , for example, the SSD of a macroblock with 4:2:0 sampling ratios is calculated as follows:

$$\mathrm{SSD}(s, c, \mathrm{MODE}|\mathrm{QP}) = \sum_{x=1,y=1}^{16,16} (S_Y[x,y] - C_Y[x,y,\mathrm{MODE}|\mathrm{QP}])^2 + \sum_{x=1,y=1}^{8,8} (S_u[x,y] - C_U[x,y,\mathrm{MODE}|\mathrm{QP}])^2 + \sum_{x=1,y=1}^{8,8} (S_V[x,y] - C_V[x,y,\mathrm{MODE}|\mathrm{QP}])^2 \quad (A.2)$$

$S_Y[x,y]$ and $C_Y[x,y,\mathrm{MODE}|\mathrm{QP}]$ represent the original and reconstructed luminance values, and S_u, S_V, C_U, C_V are the corresponding chrominance values.

The Lagrangian multiplier λ_{MODE} is given by:

$\lambda_{\mathrm{MODE,I,P}} = 0.85 \times 2^{(\mathrm{QP}-12)/3}$ for Intra mode and P slice Inter mode (A.3)

$\lambda_{\mathrm{MODE,B}} = \max(2, \min\left(4, \frac{\mathrm{QP}-12}{6}\right) \times \lambda_{\mathrm{MODE,I,P}}$ for B slice mode (A.4)

The MODE indicates a series of Intra and Inter prediction modes:

MODE ϵ {INTRA 4×4, INTRA 16×16} for I frame,

MODE ϵ {INTRA 4×4, INTRA 16×16, SKIP, 16×16, 16×8, 8×16, 8×8} for P frame,

MODE ϵ {INTRA 4×4, INTRA 16×16, DIRECT, 16×16, 16×8, 8×16, 8×8} for B frame.

where SKIP mode indicates the 16x16 mode that no motion and residual information is coded in P frame, that is, its transform coefficients are all quantised to zero. The DIRECT mode indicates no motion vector is transmitted and coded transform coefficient equal to zero in B frame.

There are two intra prediction sizes for luminance samples in H.264 standard: 4×4 block and 16×16 macroblock. The prediction samples for Intra 4×4 block are depicted in figure 2.10.

There are four Intra16×16 prediction modes and nine Intra 4×4 prediction modes in the Intra prediction:

INTRA $16 \times 16 \in$ {dc, horizontal, vertical, plane}

INTRA $4 \times 4 \in$

{dc, horizontal, vertical, diagonal_down_right, diagonal_down_left, vertical_left,

vertical_right, horizontal_up, horizontal_down}

The sub mode for each 8x8 sub-macroblock in Inter prediction is listed as follows:

INTER $8 \times 8 \in \{8 \times 8, 8 \times 4, 4 \times 8, 4 \times 4\}$

The Lagrangian cost for the 8x8 mode is the sum of each sub blocks. Then the mode with the smallest Lagrangian cost is selected as the best mode for coding the macroblock.

APPENDIX B

THE H.264 TRANSFORM, QUANTISATION, RESCALING AND INVERSE TRANSFORM PROCESS

B.1 4 ×4 Residual Transform and Quantisation in H.264:

In H.264, three integer transforms are used depending on the type of the residual data, which are depicted in figure 2.13. The H.264 transform is based on the DCT, but each row of the DCT transform is scaled and rounded to the integer. This allows the transform can be implemented using only addition and shift, without loss of decoding accuracy [2]. The differences are shown as follows:

The 4×4 two-dimensional DCT is given by:

$$Y = A \cdot X \cdot A^T = \begin{bmatrix} a & a & a & a \\ b & c & -c & -b \\ a & -a & -a & a \\ c & -b & b & -c \end{bmatrix} [X] \begin{bmatrix} a & b & a & c \\ a & c & -a & -b \\ a & -c & -a & b \\ a & -b & a & -c \end{bmatrix} \quad \text{(B.1)}$$

where:

$$a = \frac{1}{2},\ b = \sqrt{\frac{1}{2}}\cos\left(\frac{\pi}{8}\right),\ c = \sqrt{\frac{1}{2}}\cos\left(\frac{3\pi}{8}\right)$$

This matrix multiplication can be factorised to the following equivalent form:

$$Y = [C \cdot X \cdot C^T] \otimes E =$$

$$\left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & d & -d & -1 \\ 1 & -1 & -1 & 1 \\ d & -1 & 1 & -d \end{bmatrix} [X] \begin{bmatrix} 1 & 1 & 1 & d \\ 1 & d & -1 & -1 \\ 1 & -d & -1 & 1 \\ 1 & -1 & 1 & -d \end{bmatrix} \right) \otimes \begin{bmatrix} a^2 & ab & a^2 & ab \\ ab & b^2 & ab & b^2 \\ a^2 & ab & a^2 & ab \\ ab & b^2 & ab & b^2 \end{bmatrix} \quad \text{(B.2)}$$

E is a matrix of scaling factors, and d is c/b.

In order to simplify the implementation of the transform and ensure the transform remains orthogonal, d is approximated by 0.5 and b is also modified as follows:

$$a = \frac{1}{2},\ b = \sqrt{\frac{2}{5}},\ d = \frac{1}{2}$$

The 2nd and 4th row of matrix C and 2nd and 4th columns of matrix C^T are scaled by a factor of two to avoid multiplications by half which could result in loss of accuracy. The final forward transform in H.264 becomes:

$$Y = \left[C_f \cdot X \cdot C_f^T\right] \otimes E_f =$$

$$\left(\begin{bmatrix}1 & 1 & 1 & 1\\ 2 & 1 & -1 & -2\\ 1 & -1 & -1 & 1\\ 1 & -1 & 2 & -1\end{bmatrix}[X]\begin{bmatrix}1 & 2 & 1 & 1\\ 1 & 1 & -1 & -2\\ 1 & -1 & -1 & 2\\ 1 & -2 & 1 & -1\end{bmatrix}\right) \otimes \begin{bmatrix}a^2 & \frac{ab}{2} & a^2 & \frac{ab}{2}\\ \frac{ab}{2} & \frac{b^2}{4} & \frac{ab}{2} & \frac{b^2}{4}\\ a^2 & \frac{b^2}{4} & a^2 & \frac{b^2}{4}\\ \frac{ab}{2} & \frac{b^2}{4} & \frac{ab}{2} & \frac{b^2}{4}\end{bmatrix} \quad \text{(B.3)}$$

The inverse transform in H.264 is defined as follows:

$$Y = C_i^T \cdot (Y \otimes E_i) \cdot C_i = \begin{bmatrix}1 & 1 & 1 & \frac{1}{2}\\ 1 & \frac{1}{2} & -1 & -1\\ 1 & -\frac{1}{2} & -1 & 1\\ 1 & -1 & 1 & -\frac{1}{2}\end{bmatrix}\left([X] \otimes \begin{bmatrix}a^2 & ab & a^2 & ab\\ ab & b^2 & ab & b^2\\ a^2 & ab & a^2 & ab\\ ab & b^2 & ab & b^2\end{bmatrix}\right) \quad \text{(B.4)}$$

Quantisation

The basic forward quantiser (Equation 2.9) operation is:

$$Z_{ij} = round\left(\frac{Y_{ij}}{Qstep}\right) \quad \text{(B.5)}$$

Y_{ij} is a coefficient of the transform described above, $Qstep$ is a quantiser step size and Z_{ij} is a quantised coefficient. A total of 52 values of $Qstep$ are supported by the standard, $Qstep$ doubles in size for every increment of 6 in QP. The relationship between QP and $Qstep$ is described as follows [3]:

QP	0	1	2	3	4	5	6	…	51
$Qstep$	0.625	0.702	0.787	0.884	0.992	1.114	1.250	…	224

Table B-1 Quantisation step sizes in H.264 CODEC

The ration between successive $Qstep$ values is chosen to be $\sqrt[6]{2} = 1.2246$, and any value of $Qstep$ can be derived from the first 6 values in the table ($QP0 - QP5$) as follows:

$$Qstep(QP) = Qstep(QP\%6) \cdot 2^{floor(QP/6)} \quad \text{(B.6)}$$

In order to avoid division and/or floating point arithmetic, the post-scaling matrix E_f is incorporated into the forward quantised process.

$$Z_{ij} = round\left(\left[\mathrm{C_f} \cdot \mathrm{X} \cdot \mathrm{C_f^T}\right] \cdot \frac{PF}{Qstep}\right) \quad \text{(B.7)}$$

where PF is the $\mathrm{a}^2,\ \mathrm{ab}/2,\ \text{or}\ \mathrm{b}^2/4$ in the post-scaling matrix $\mathrm{E_f}$. In order to avoid any division operations, the factor $(PF/\,Qstep)$ is implemented in the reference model software [4] as a multiplication by a factor MF and a right shift.

$$Z_{ij} = round\left(\left[\mathrm{C_f} \cdot \mathrm{X} \cdot \mathrm{C_f^T}\right] \cdot \frac{MF}{2^{qbits}}\right) \quad \text{(B.8)}$$

where $\frac{MF}{2^{qbits}} = \frac{PF}{Qstep}$, and $\mathrm{qbits} = 15 + \mathrm{floor}(\mathrm{QP}/6)$. Then the first six values of MF used by the H.264 reference software encoder are given as follows:

QP	MF positions (0,0), (2,0), (2,2), (0,2)	MF positions (1,1), (1,3), (3,1), (3,3)	MF positions (Other positions)
0	13107	5243	8066
1	11916	4660	7490
2	10082	4194	6554
3	9362	3647	5825
4	8192	3355	5243
5	7282	2893	4559

Table B-2 Multiplication factor MF

In integer arithmetic, the equation (B8) can be implemented as follow:

$$|Z_{ij}| = \left(|W_{ij}| \cdot \mathrm{MF} + \mathrm{f}\right) \gg \mathrm{qbits}$$

$$\mathrm{sign}\left(Z_{ij}\right) = sign(W_{ij}) \quad \text{(B.9)}$$

where $\mathrm{W_{ij}} = \mathrm{C_f} \cdot \mathrm{X} \cdot \mathrm{C_f^T}$, f is $2^{qbits}/3$ for intra block or $2^{qbits}/6$ for inter blocks that defined in the reference model software.

The rescaling (inverse quantisation) and transform

Due to the quantisation stage is not reversible and so a more accurate term rescale is often used in the inverse process. The basic scaling operation (Equation 2.9) is:

$$Y'_{ij} = Z_{ij} \times Qstep \quad \text{(B.10)}$$

The pre-scaling factor (matrix E_i) for the inverse transform is incorporated in this operation, and a constant scaling factor of 64 is used to avoid rounding errors:

$$W'_{ij} = Z_{ij} \cdot Qstep \cdot PF \cdot 64 \quad \text{(B.11)}$$

W'_{ij} is the scaled coefficient which is transformed by the core inverse transform $C_i^T \cdot W \cdot C_i$ (Equation B4). The parameter $V = Qstep \cdot PF \cdot 64$ is defined in H.264 for $QP0 - QP5$.

QP	V positions (0,0), (2,0), (2,2), (0,2)	V positions (1,1), (1,3), (3,1), (3,3)	V positions (Other positions)
0	10	16	13
1	11	18	14
2	13	20	16
3	14	23	18
4	16	25	20
5	18	29	23

Table B-3 Scaling factor V

For larger values of $QP > 5$, each coefficient is increased by a factor of two for every increment of six in QP. Then the scaling operation becomes:

$$W'_{ij} = Z_{ij} \cdot V_{ij} \cdot 2^{\text{floor}(QP/6)} \quad \text{(B.12)}$$

The complete inverse transform and scaling process becomes:

$$Y'_{ij} = \text{round}\,(C_i^T \cdot W'_{ij} \cdot C_i \cdot \frac{1}{64}) \quad \text{(B.13)}$$

B.2 4 × 4 Luma DC Coefficient Transform and Quantisation (16 × 16 Intra-mode)

If the macroblock is encoded in 16×16 intra prediction mode, the Hadamard transform H_2 (Figure 2.13) is used in addition to the transform described above:

$$Y_D = \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} [W_D] \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \right) /2 \quad \text{(B.14)}$$

where $W_D (= C_f \cdot X \cdot C_f^T)$ is the block of 4×4 DC coefficients. Then the output coefficients $Y_{D(i,j)}$ are quantised as following:

$$|Z_{D(i,j)}| = \left(|Y_{D(i,j)}| \cdot MF_{(0,0)} + 2f\right) \gg (qbits + 1)$$

$$\text{sign}\left(Z_{D(i,j)}\right) = sign(Y_{D(i,j)}) \quad \text{(B.15)}$$

where $MF_{(0,0)}$ is the scaling factor MF for position (0,0) in Table B-2, and f are defined same as above.

The rescaling and transform

The inverse Hadamard transform is:

$$W_{QD} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} [Z_D] \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix} \right) \quad \text{(B.16)}$$

The rescale operation is:

$$W'_{D(i,j)} = W_{QD(i,j)} V_{(0,0)} 2^{\text{floor}(QP/6)-2} \quad (QP \geq 12)$$

$$W'_{D(i,j)} = [W_{QD(i,j)} V_{(0,0)} + 2^{1-\text{floor}(QP/6)}] \gg (2 - \text{floor}(QP/6)) \quad (QP < 12) \quad \text{(B.17)}$$

where $V_{(0,0)}$ is the scaling factor V for position (0,0) in Table B-3. Then the rescaled DC coefficients $W'_{D(i,j)}$ is reverse transformed($C_i^T \cdot W'_{D(i,j)} \cdot C_i$). The additional Hadamard transform is to make sure the energy is concentrated further into a small number of significant coefficients.

B.3 2 × 2 Chroma DC Coefficient Transform and Quantisation

Each 4×4 block in the chrominance components is transformed by using H_1 first. Then the DC coefficients of each 4×4 block of chrominance coefficients are grouped in a 2×2 block ($W_{D)}$, and the third transform H_3 is used to transform the 4 DC coefficients of each chrominance component:

$$W_{QD} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} [W_D] \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad \text{(B.18)}$$

The quantisation operation is:

$$|Z_{D(i,j)}| = \left(|Y_{D(i,j)}| \cdot \mathrm{MF}_{(0,0)} + 2\mathrm{f}\right) \gg (\mathrm{qbits} + 1)$$

$$\mathrm{sign}\left(Z_{D(i,j)}\right) = sign(Y_{D(i,j)}) \quad \text{(B.19)}$$

The inverse transform is:

$$W_{QD} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} [Z_D] \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad \text{(B.20)}$$

The inverse transform is followed by the scaling:

$$W'_{D(i,j)} = W_{QD(i,j)} V_{(0,0)} 2^{\mathrm{floor}(QP/6)-1} \ (QP \geq 6)$$

$$W'_{D(i,j)} = [W_{QD(i,j)} V_{(0,0)}] \gg 1 \ (QP < 6) \quad \text{(B.21)}$$

Then the rescaled coefficients $W'_{D(i,j)}$ are reverse transformed by using $C_i^T \cdot W'_{D(i,j)} \cdot C_i$.

[1] K. P. Lim, G. Sullivan, and T. Wiegand “Text Description of Joint Model Reference Encoding Methods and Decoding Concealment Methods.” Document JVT-N046, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, Hong Kong, Jan. 2005.

[2] Iain E.G. Richardson, H.264 and MPEG-4 Video Compression, Thomson Press (India) Ltd., New Delhi, England. P187 (2003).

[3] Iain E.G. Richardson, 4×4 Transform and Quantization in H.264/AVC, VCodex Ltd White Paper, April 2009, http:// www.vcodex.com.

[4] H.264 Reference Software Version JM6.1d, March 2003

APPENDIX C

POST-ENCODING STAGE OF RATE CONTROL IN SCALABLE VIDEO CODING

The frame layer rate control scheme consists of two stages: pre-encoding and post-encoding. The pre-encoding stage is described in section 4.3.4 Rate Control Scheme. The objective of the pre-encoding stage is to compute quantisation parameter for all frames. There are three tasks in post-encoding stage: update the parameters a_1 and a_2 of the linear prediction model (Equation 4.3), the parameters of Rate-Quantization model (Equation 4.1), and determine the number of frames needed to be skipped [1]. The linear and quadratic models parameters are updated through a linear regression method that similar to MPEG-4 Q2 [2][3][4].

1) Quadratic R-D model update:

The quantisation parameter QP_i for the current frame i is computed based on the Q2 model:

$$R_i = X_1 \cdot \frac{MAD_i}{QP_i} + X_2 \cdot \frac{MAD_i}{QP_i^2} \quad \text{(C.1)}$$

where $R_i = (R - H)$ denotes the total number of bits assigned for the current frame i excluding bits used for header and motion vectors, and MAD_i denote mean absolute difference between the reference frame and the current frame i.

Hence,

$$X_1 = \frac{R_i - X_2 \times MAD_i \times QP_i^{-2}}{MAD_i \times QP_i^{-1}} = \frac{R_i \times QP_i}{MAD_i} - \frac{X_2}{QP_i}$$

$$X_2 = \frac{R_i - X_1 \times MAD_i \times QP_i^{-1}}{MAD_i \times QP_i^{-2}} = \frac{R_i}{MAD_i \times QP_i^{-2}} - X_1 \times QP_i \quad \text{(C.2)}$$

The modelling parameters X_1 and X_2 are solved by using the least-square method based on the previous data [5].

$$X_1 = \frac{\sum_{i=1}^{n} \frac{QP_i \times R_i}{MAD_i} - X_2 \times \sum_{i=1}^{n} QP_i^{-1}}{n}$$

$$X_2 = \frac{n \sum_{i=1}^{n} \frac{R_i}{MAD_i} - \left(\sum_{i=1}^{n} QP_i^{-1}\right)\left(\sum_{i=1}^{n} \frac{QP_i \times R_i}{MAD_i}\right)}{n \sum_{i=1}^{n} QP_i^{-2} - \sum_{i=1}^{n} QP_i^{-2}} \quad \text{(C.3)}$$

where n is the number of selected past frames. The selected past frames are updated for each step for better modelling. A sliding window mechanism is provided in [2] [6] to

adaptively smooth the impact of a scene change for certain number of frames in updating the quadratic R-D model. If the complexity of the current frame is increased, such as high motion scenes, then a small number of data points with recent data are selected. Otherwise, a larger number of data points with previous data are selected.

Mathematically, let

$$\delta(n) = \min\{\frac{MAD_{t-1}}{MAD_t}, \frac{MAD_t}{MAD_{t-1}}\} \quad \text{(C.4)}$$

where t is the time instant of coding, the size of the sliding window for the current frame is given by

$$W_s^p(n) = \min\{W_s^p(n-1) + 1, \delta(n) \times Max\_Sliding\_window\} \quad \text{(C.5)}$$

where $Max\_Sliding\_window$ is a preset constant [2].

2) The linear MAD prediction model

The initial values of a_1 and a_2 are set to 1 and 0, respectively. They are updated by a linear regression method that similar to the quadratic R-D model.

3) Frame-Skipping Control

The objective of the frame-skipping control is to prevent buffer overflow. After encoding each current frame, the generated bits is added into the buffer, if the current buffer status is above 80 percents, the encoder will skip the upcoming frames.

[1] Z. G. Li, F. Pan, K. P. Lim, G. Feng, X. Lin, and S. Rahardja, "Adaptive basic unit layer rate control for JVT," presented at the 7th JVT Meeting, Pattaya II JVT-G012-rl Thailand, Mar. 2003.

[2] H.J.Lee and T.H.Chiang and Y.Q.Zhang. Scalable Rate Control for MPEG-4 Video. IEEE Trans. Circuit Syst. Video Technology, 10: 878-894, 2000.

[3] A.Vetro, H.Sun and Y.Wang. MPEG-4 rate control for multiple video objects. IEEE Trans. Circuit Syst. Video Technology, 9: 186-199, 1999.

[4] T. Chiang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate distortion model," IEEE Trans. Circuits Syst. Video Technol., vol. 7, pp. 246–250, Feb. 1997.

[5] D. Wu et al., "On end-to-end architecture for transporting MPEG-4 video over the internet," IEEE Trans. Circuits Syst. Video Technol., vol. 10, pp. 923–941, Sept. 2000.

[6] Z. G. Li, Lin Xiao, C. Zhu and Pan Feng. A Novel Rate Control Scheme for Video Over the Internet. In Proceedings ICASSP 2002, Florida, USA, 2065--2068, May 13-17, 2002.

APPENDIX D

PUBLICATIONS

1. Guang Y. Zhang, Abdelrahman Abdelazim, Stephen James Mein, Martin Roy Varley and Djamel Ait-Boudaoud, “A simplified rate control algorithm for H.264/SVC”, SPIE conference, April 2011.

2. Guang Y. Zhang, Abdelrahman Abdelazim, Stephen James Mein, Martin Roy Varley, and Djamel Ait-Boudaoud, “Inter-view Reference Frame Selection in Multi-view Video Coding”, Data Compression Conference, Snowbird, Utah, March 20 - 22, 2013. (Accepted as a poster)

A Simplified Rate Control Algorithm for H.264/SVC

Guang Y. Zhang[a], Abdelrahman Abdelazim[a], Stephen James Mein[a], Martin Roy Varley[a]
and Djamel Ait-Boudaoud[b].
[a] Applied Digital Signal and Image Processing Research Centre, School of Computing, Engineering and Physical Sciences, University of Central Lancashire, UK
{AAbdelazim, GYZhang1 SJMein, MRVarley}@ uclan.ac.uk
[b] University of Portsmouth. UK, djamel.ait-boudaoud@port.ac.uk

ABSTRACT

The objective of scalable video coding is to enable the generation of a unique bitstream that can adapt to various bit-rates, transmission channels and display capabilities. The scalability is categorised in terms of temporal, spatial, and quality. Effective Rate Control (RC) has important ramifications for coding efficiency, and also channel bandwidth and buffer constraints in real-time communication.

The main target of RC is to reduce the disparity between the actual and target bit-rates. In order to meet the target bit-rate, a predicted Mean of Absolute Difference (MAD) between frames is used in a rate-quantisation model to obtain the Quantisation Parameter (QP) for encoding the current frame.

The encoding process exploits the interdependencies between video frames; therefore the MAD does not change abruptly unless the scene changes significantly. After the scene change, the MAD will maintain a stable slow increase or decrease. Based on this observation, we developed a simplified RC algorithm. The scheme is divided in two steps; firstly, we predict scene changes, secondly, in order to suppress the visual quality, we limit the change in QP value between two frames to an adaptive range. This limits the need to use the rate-quantisation model to those situations where the scene changes significantly.

To assess the proposed algorithm, comprehensive experiments were conducted. The experimental results show that the proposed algorithm significantly reduces encoding time whilst maintaining similar rate distortion performance, compared to both the H.264/SVC reference software and recently reported work.

Keywords: Scalable, Rate Control, H.264, Inter-View, SVC, Sub-Pixel

1. INTRODUCTION

Nowadays the amount of digital video applications is rapidly increasing. Videos are being watched at home, on the internet, using mobile devices and portable DVD players. All those devices have different capabilities and use different types of transmission systems. The object of scalable video coding is to enable a video stream to adapt to variety of devices with different capabilities. The main objective of Scalable Video Coding (SVC) is to achieve the scalability without a significant loss in coding efficiency.

A rate control algorithm can adjust the encoder parameters to produce a higher quality decoded frame with the constraints of decoder buffer and network bandwidth. One of the most important parameter is the Quantisation Parameter (QP), which regulates how much spatial detail is saved. As the QP is increased the detail will be aggregated so that the bit rate drops, the quality of the frame will be decreased. Due to the important role of the QP in regulating the bit rate and the Rate Distortion Optimisation (RDO), and due to the fact that no information about the scene is available before encoding the following 'chicken and egg' dilemma needed to be solved. The QP should be first determined in order to perform the RDO by using the mean absolute difference (MAD) of the current frame or macroblock (MB) [1, 2], however, the MAD of the current frame or MB is calculated after the RDO. Moreover, the available channel bandwidth can either be constant bit rate (CBR) or variable bit rate (VBR), thus both of them need to be considered.

Due to the importance of the rate control in video coding, a number of solutions [3, 5] have been proposed to solve this problem. A solution based on the concepts of a basic unit and a linear mode was proposed in [3]. In this, a linear model is used to predict the MAD of the current basic unit of the co-located position of the previous frame; the basic unit can be a frame, a slice, or a MB. The corresponding QP is calculated based on a quadratic rate-distortion (R-D) model [1, 2], then the QP is used for the RDO for each MB in the current basic unit. The real MAD of the current MB is obtained after the RDO is performed, then used to predict the MAD for the next basic unit using the same linear model. The quadratic rate-distortion model is the core of the algorithm, the relationship between QP, bit rate and MAD is illustrated in this model. The demanded QP is calculated based on this model. Available channel bandwidth is always changing; therefore in most applications a target bit rate needs to be calculated first. A method to compute the target bit rate for the current frame was proposed in [4], it exploits the fluid traffic model and linear tracking theory. The target bit rate is established from the number of bits remaining in the buffer and the number of bits used for coding the previous frame, it is also updated on frame by frame basis. In this algorithm, the header bits are excluded for the target bits, thus the target bit means the texture bits or the residual bits.

The proposed rate control, above, is composed of two layers: group of picture (GOP) layer and frame layer, if the basic unit is selected as a frame, otherwise, the basic unit layer rate control should be added. By using this scheme the chicken and egg dilemma is solved and it is applicable in both the VBR and CBR case. The disadvantage of the linear prediction model is that it is not suitable for predicting the abrupt MAD fluctuation. If the complexity of the scene is changed suddenly the MAD will not be predicted accurately by using the temporal linear model.

A switched MAD prediction scheme is introduced in [5] to solve the problem of the scene change abrupt. The proposed MAD prediction scheme is switched between a temporal prediction model and a spatial prediction model depending on their difference from the actual MAD of the previous frame. If the temporal prediction model absolute difference is smaller than the spatial prediction model, then the temporal prediction model is chosen, otherwise the spatial prediction model is chosen. The experiment results show that the switched scheme is more accurate to predict the MAD. In order to reduce the coding complexity, a linear R-Q model is also proposed in the literature.

In this paper, we propose an algorithm where the QP can be obtained directly from the predicted MAD. The algorithm is divided in two steps; firstly, scene changes are predicted, secondly, in order to suppress the visual quality, the change in QP value between two frames is limited to an adaptive range.

The rest of this paper is organised as follows. Section 2 gives a brief overview of the SVC. In section 3, the proposed algorithm is presented. Section 4 presents the experimental results. Finally, the conclusion and future work are given in section 5.

2. OVERVIEW OF THE SCALABLE VIDEO CODING SVC

2.1. Rate Distortion Optimization (RDO)

Similar to H.264/AVC, the motion estimation and mode decision process in SVC is performed by minimizing the rate-distortion cost function below:

$$J = Distortion + \lambda_{MODE} \cdot Rate \tag{6}$$

The *Distortion* measurement quantifies the quality of the reconstructed pictures whilst *Rate* measures the bits needed to code the macroblock using the particular mode. In the first stage of motion prediction, an integer-pixel motion search is performed for each 16×16 square MB of the frame to find a best match within a search range in the reference frames. After this, the values at half-pixel positions and quarter-pixel positions around the best match are searched to find an even better match. This process is repeated for the eight enclosed partitions of the MB (two 16×8, two 8×16, and four 8×8), and the mode that minimizes a cost function is selected as the best encoding mode. Furthermore, if the four 8×8 blocks are selected as the best mode the search is repeated for the eight enclosed sub-MBs (two 8×4, two 4×8, and four 4×4). In addition, the skip mode, direct mode and intra modes are also supported.

2.2. Temporal Scalability

Temporal scalability is provided when the bit stream that is obtained by removing all temporal layers with a temporal identifier greater than a natural number produces a valid bit stream for a given decoder. This means all the picture is coded with defined temporal dependencies. The enhancement layer pictures are typically coded as B-pictures, where the reference picture lists 0 and 1 are restricted to the temporally preceding and succeeding picture, respectively, with a temporal layer identifier less than the temporal layer identifier of the predicted picture.

2.3. Spatial Scalability

For supporting spatial Scalability, in the SVC, each layer corresponds to a supported spatial resolution and is referred to by a spatial layer. In each spatial layer, motion-compensated prediction and intra-prediction are employed similar to the processes used for the base layer. Additionally an inter-layer prediction mechanism is incorporated.

2.4. Quality Scalability

Quality scalability is considered as a special case of spatial scalability with identical picture sizes for base and enhancement layers, therefore the same motion-compensated predictions method as the spatial scalability including the inter-layer prediction are employed.

2.5. Group of Pictures (GOP)

In H.264/SVC a Group of Pictures (GOP) is an encoding of a contiguous subset of frames from a video sequence; each GOP consists of all frames between two successive frames at the lowest temporal resolution, plus the second of the temporal base layer frames (the first is considered to belong to the previous GOP). All information required to decode any one frame from the GOP is contained within it.

3. PROPOSED ALGORITHM

The quadratic rate control scheme has been proposed to solve the dilemma. However, the QP calculation in the quadratic model for each frame or MB significantly increases the coding complexity. In order to decrease the coding complexity, it is necessary to simplify the model and keep the same accuracy at any time. In the video codec, the coding exploits the interdependencies of video frames [7]. However, in many video sequences, frequently, the scenes change suddenly. Furthermore, in order to suppress the visual quality "beating" or "pulsing" [8], the QP cannot be increased too fast. It can be regarded that the QP is unchanged or changing slowly before the scene is suddenly changed. Two examples of the relationship between MAD and QP are shown in Fig 1—Fig 4. The figures show that the QP changes with the MAD. When the MAD is increased the QP is also increased. According to the above consideration, we propose the following three steps algorithm to adaptively obtain QP:

- First, an initial QP will be used to encode the frame or MB;
- Second, if the predicted MAD is greater than the previous MAD, the QP is directly increased by a Threshold (TH);
- Otherwise, a quadratic model will be used to decide a proper QP for the current frame or MB.

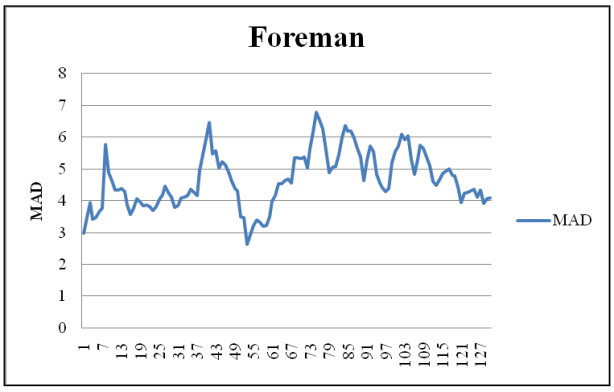


Fig. 1. The prdicted MAD for the current frame. Analysis of the P-frames from FOREMAN@QCIF-15 Hz .

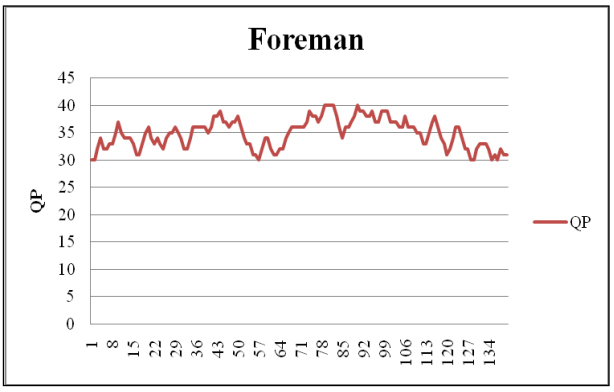


Fig. 2. The actual QP for the corresponding frame. Analiysis of the P-frames from FOREMAN@QCIF-15 Hz .

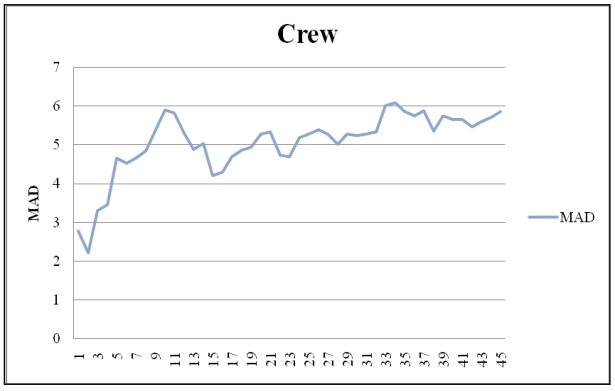


Fig. 3. The prdicted MAD for the current frame. Analysis of the P-frames from CREW@QCIF--15 Hz .

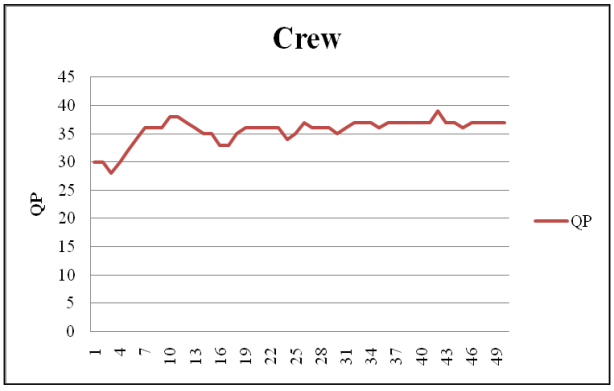


Fig. 4. The actual QP for the corresponding frame. Analysis of the P-frames from CREW@QCIF-15 Hz .

The algorithm is summarised using the following flowchart:

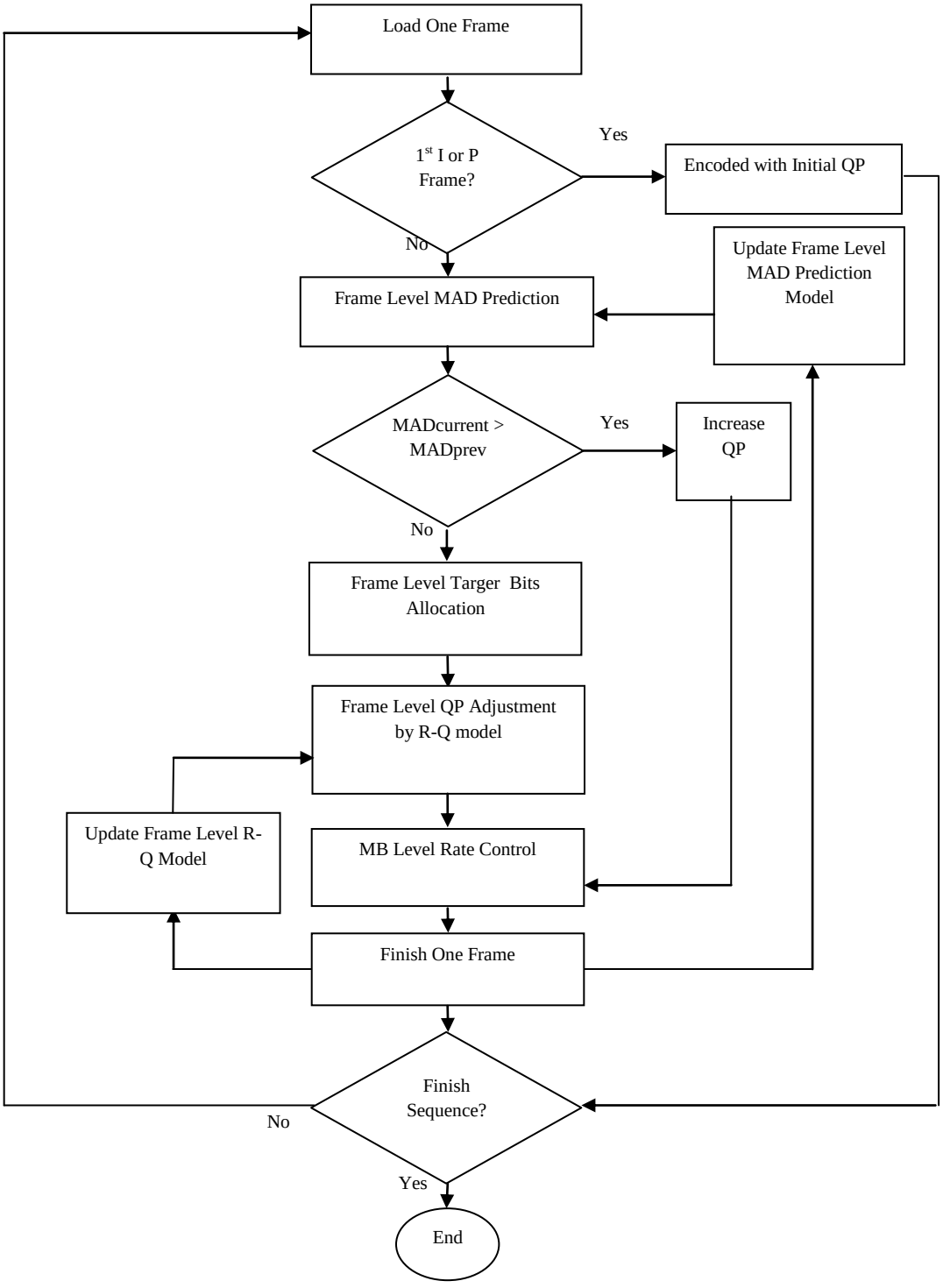


Fig. 5. Block diagram of the proposed rate control scheme at frame level

4. EXPERIMENTAL RESULTS

In order to evaluate the proposed algorithm, a comprehensive set of experiments have been carried out. Recently, a new version of Joint Scalable Video Model (JSVM) software is updated for the scalable video coding. The scheme is implemented on the latest JSVM 9.19.9 encoder [9]. The test platform uses an Intel Core 2 CPU 6420 @ 3.20GHz with 3.0 GB RAM. The Intel VTune performance analyzer was used to measure the number of machine cycles differences which reflects the total encoding Time Saving. Additionally, bit rate and PSNR have been used to evaluate the proposed algorithm performance against the JSVM encoder.

Sequence	*RC*	Act. Bit (kb/s)	PSNR (dB)	Time Saving (%)
Foreman	Standard	65.91	32.19	+4.33
	Ours	50.11	30.25	
Bus	Standard	132.92	29.81	+1.40
	Ours	85.28	27.30	
Crew	Standard	96.78	32.36	+4.42
	Ours	67.91	30.64	
Soccer	Standard	53.50	29.91	+1.92
	Ours	53.49	29.85	
City	Standard	51.86	31.19	+4.19
	Ours	32.88	28.61	

Table 1. Comparison between the Proposed Algorithm and the JSVM 9.19.9 software.

Several standard video sequences were used in the experiments as shown in the table. From the table it can be seen that the proposed algorithm achieves a significant reduction in the bit rate with a acceptable reduction in PSNR. Additionally a time saving of up to 4% has been accomplished when compared to the standard. The above result is shown for temporal scalability.

5. CONCLUSION AND FUTURE WORK

In this paper, the major problem of complexity in the scalable video coding is addressed. We proposed a direct way to obtain the QP based on the MAD change. From the simulation results, the proposed approach always outperforms the standard and the time saving is up to 4%.

Two additional steps need to be added to the algorithm; firstly, a buffer control needs to be added to enable the algorithm to achieve a target bitrate when required. Secondly, the deployment should be extended to spatial and coarse-grain scalabilities to limit the need of using the rate-quantisation model to those situations where the scene changes significantly in the base layer.

REFERENCES

[1] H.J. Lee and T.H. Chiang and Y.Q. Zhang, "Scalable Rate Control for MPEG-4 Video", IEEE Trans. Circuit Syst. Video Technology, papers 10, (878-894) 2000.

[2] A.Vetro, H.Sun and Y.Wang, "MPEG-4 rate control for multiple video objects", IEEE Trans. Circuit Syst. Video Technology, papers 9, (186-199)1999.

[3] Z. G. Li, F. Pan, K. P. Lim, G. N. Feng, X. Lin and S. Rahardaj, "Adaptive basic unit layer rate control for JVT, JVT-G012", Pattaya II, Thailand, (7-14)2003.

[4] Z. G. Li, Lin Xiao, C. Zhu and Pan Feng, "A Novel Rate Control Scheme for Video Over the Internet", In Proceedings ICASSP 2002, Florida, USA, (2065—2068) 2002.

[5] Yang Liu, Zhengguo G. Li, and Yeng Chai Soh., "A Novel Rate Control Scheme for Low Delay Video Communication of H.264/AVC Standard", IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 17, NO. 1,2007.

[6] Wiegand, T., Sullivan, G., J., Bjontegaard, G. and Luthra, A., "Overview of the H.264/AVC video coding standard," IEEE Trans. Circuits Syst. Video Technol., 13(7), 560–576(2003).

[7] Iain E.G. Richardson, Video Codec Design, Thomson Press (India) Ltd., New Delhi, England. P41-42 (2002).

[8] Athanasios Leontaris, Alexis Michael Tourapis, Dolby Laboratories, "Rate Control reorganization in the Joint Model (JM) reference software," Joint Video Team of ISO/IEC MPEG and ITU-T VCEG, JVT-W042, April, 2007.

[9] JSVM 9.19.9 Software Package, CVS server for the JSVM software, April.2010.

Inter-view Reference Frame Selection in Multi-view Video Coding

Guang Y. Zhang[a], Abdelrahman Abdelazim[a], Stephen James Mein[a], Martin Roy Varley[a] and Djamel Ait-Boudaoud[b].

[a] Applied Digital Signal and Image Processing Research Centre, School of Computing, Engineering and Physical Sciences, University of Central Lancashire, UK

{GYZhang1, AAbdelazim, SJMein, MRVarley}@ uclan.ac.uk

[b] University of Portsmouth. UK, djamel.ait-boudaoud@port.ac.uk

Abstract

Multiple video cameras are used to capture the same scene simultaneously to acquire the multiview view coding data; obviously, over-large data will affect the coding efficiency. Due to the video data is acquired from the same scene, the inter-view similarities between adjacent camera views are exploited for efficient compression. Generally, the same objects with different viewpoints are shown on adjacent views. On the other hand, containing objects at different depth planes, and therefore perfect correlation over the entire image area will never occur. Additionally, the scene complexity and the differences in brightness and color between the video of the individual cameras will also affect the current block to find its best match in the inter-view reference picture. Consequently, the temporal-view reference picture is referred more frequently. In order to gain the compression efficiency, it is a core part to disable the unnecessary inter-view reference. The idea of this paper is to exploit the phase correlation to estimate the dependencies between the inter-view reference and the current picture. If the two frames with low correlation, the inter-view reference frame will be disabled. In addition, this approach works only on non-anchor pictures. Experimental results show that the proposed algorithm can save 16% computational complexity on average, with negligible loss of quality and bit rate. The phase correlation process only takes up 0.1% of the whole process.

Index Terms—inter-view reference frame selection, phase correlation, multi-view video coding (MVC).

I. Introduction

MVC has been developed as an extension of the ITU-T Recommendation H.264 | ISO/IEC International Standard ISO/IEC 14496-10 advanced video [1]. MVC provides an efficient video coding scheme for multiple views of a video scene captured from different synchronized video cameras. Stereo-paired video for 3-D viewing and Free Viewpoint Television are the main targeted applications of MVC [2]. However, these applications require a large storage and high transmission rate. A lot of coding schemes have been developed to improve the coding efficiency, the one proposed by HHI [3] performs better than the simulcast anchors, it uses the hierarchical B prediction structure for temporal view and inter-view prediction is also applied to every view. The variable block size prediction techniques of H.264 are used in both temporal and inter-view prediction, however, the sophisticated motion estimation process significantly increase the computational complexity . In order to overcome the uniform search strategy, the correlation between views inherent is taken into account in [4]. The mode complexity and motion homogeneity of a macroblock (MB) in the current view are analyzed based on the previously coded neighbour views at the co-located position; furthermore, the global disparity vector is used to locate the corresponding position. Based on the two conditions, the candidate mode and search range of motion estimation is determined, in addition, the disparity search is selectively enabled. A fast inter frame prediction algorithm is proposed in [55] to reduce the computational complexity. The prediction direction is decided according to the co-located MB in the temporal view and the search range in the view direction is decided by the displacement of the two cameras. In [6], the region homogeneity is determined by calculating the difference between the coding MB and the corresponding MB in the reference block. However, the objects at different depth planes and light condition are not taken into account when determine the region homogeneity in these approaches, meanwhile, the additional information of the camera geometry is required to decide the co-located position. These proposed algorithms provide a good coding efficiency compared with the standard coding. However, these algorithms increase the complexity on determining the search range, candidate mode and region homogeneity. Though inter-view reference frame can be used to reduce the computational complexity, it is not always worked for all the video sequence and every frame in the sequence. Additionally, the disparity search should be skipped if the correlation between the current MB and reference MB is small. Moreover, according to our experimental

result, it shows that only 10.2% MBs in view 1, view 3 and view 5 of the ballroom sequence found their best match in the view direction and 5.5% in the exit sequence. Based on these observations, the two frames in the inter-view and current view are divided into N blocks of X by Y pels, and the phase correlation is employed to obtain the phase correlation between these co-located blocks, and then the inter-view reference frame is selectively disabled based on the number of blocks with less correlation.
The rest of this paper is organized as follows. Section II gives an introduction on phase correlation and the proposed algorithm. Section III presents the experimental results and discussion. Finally, the conclusion is given in sections IV.

II. Phase correlation and proposed algorithm

In [7], an elegant method called phase correlation is proposed to align two images which have been shifted relative to each other. This method measures the movement between the two fields directly from the phase. In here, the phase correlation method is used to determine the similarities between the current frame and the translated inter-view reference frame that was captured at the same time. The correlation basic principles are briefed as below.
Assuming a translational shift between the two frames

$$s_{k+1}(n_1, n_2) = s_k(n_1 + d_1, n_2 + d_2) \quad \text{Eq. 1}$$

Their 2-D Fourier transforms are

$$s_{k+1}(f_1, f_2) = s_k(f_1, f_2)\exp[j2\pi(d_1 f_1 + d_2 f_2)] \quad \text{Eq. 2}$$

Therefore the shift in the spatial-domain is reflected as a phase change in the spectrum domain. The cross-correlation between the two frames is

$$c_{k,k+1}(n_1, n_2) = s_{k+1}(n_1, n_2) * s_k^*(-n_1, -n_2) \quad \text{Eq.3}$$

whose Fourier transform is

$$C_{k,k+1}(f_1, f_2) = s_{k+1}(f_1, f_2).\, s_k^*(f_1, f_2) \quad \text{Eq. 4}$$

The phase is obtained by normalizing the cross-power spectrum by its magnitude

$$\Phi\left[C_{k,k+1}(f_1, f_2)\right] = \frac{s_{k+1}^*(f_1,f_2).s_k(f_1,f_2)}{|s_{k+1}^*(f_1,f_2).s_k(f_1,f_2)|} \quad \text{Eq. 5}$$

By equation 2 and 5, we have

$$\Phi\left[C_{k,k+1}(f_1, f_2)\right] = \exp\left[-j2\pi(d_1 f_1 + d_2 f_2)\right] \quad \text{Eq. 6}$$

2-D inverse transform is given by

$$c_{k,k+1}(n_1, n_2) = \delta(n_1 - d_1, n_2 - d_2) \quad \text{Eq. 7}$$

Then the motion vector is obtained by using the location of the pulse. The pulse is located by finding the highest peak. Due to the perfect correlation over the entire image area will never occur, the sub-images with small sections of the image will be taken into consideration, for which there will be points of strong correlation [8]. In implementation, both the current and inter-view reference frame is divided into 32×32 blocks. The phase correlation is calculated between the current MB and corresponding MB at the co-located position in the inter-view reference without considering the geometry of the cameras. The proposed inter-view reference skip decision strategy is sketched in Fig.1. From this figure, the inter-view reference skip decision is described in detail. If phase correlation between the two MBs is bigger than a threshold, TH_1, the MB is considered to be with high correlation with the MB in the inter-view reference, and then the number Count will be increased by one, otherwise the Count will keep the same. If Count is larger than a threshold, TH_2, the inter-view reference is considered with high correlation with the current frame, and then it is enabled in the current reference list, otherwise, it is disabled in the reference list. Based on the observation from experimental results, TH_1 and TH_2 are set to 0.5 and 100 respectively, fixed for each QP level and sequences with size 640×480, the TH_2 can be changed based on the frame size.

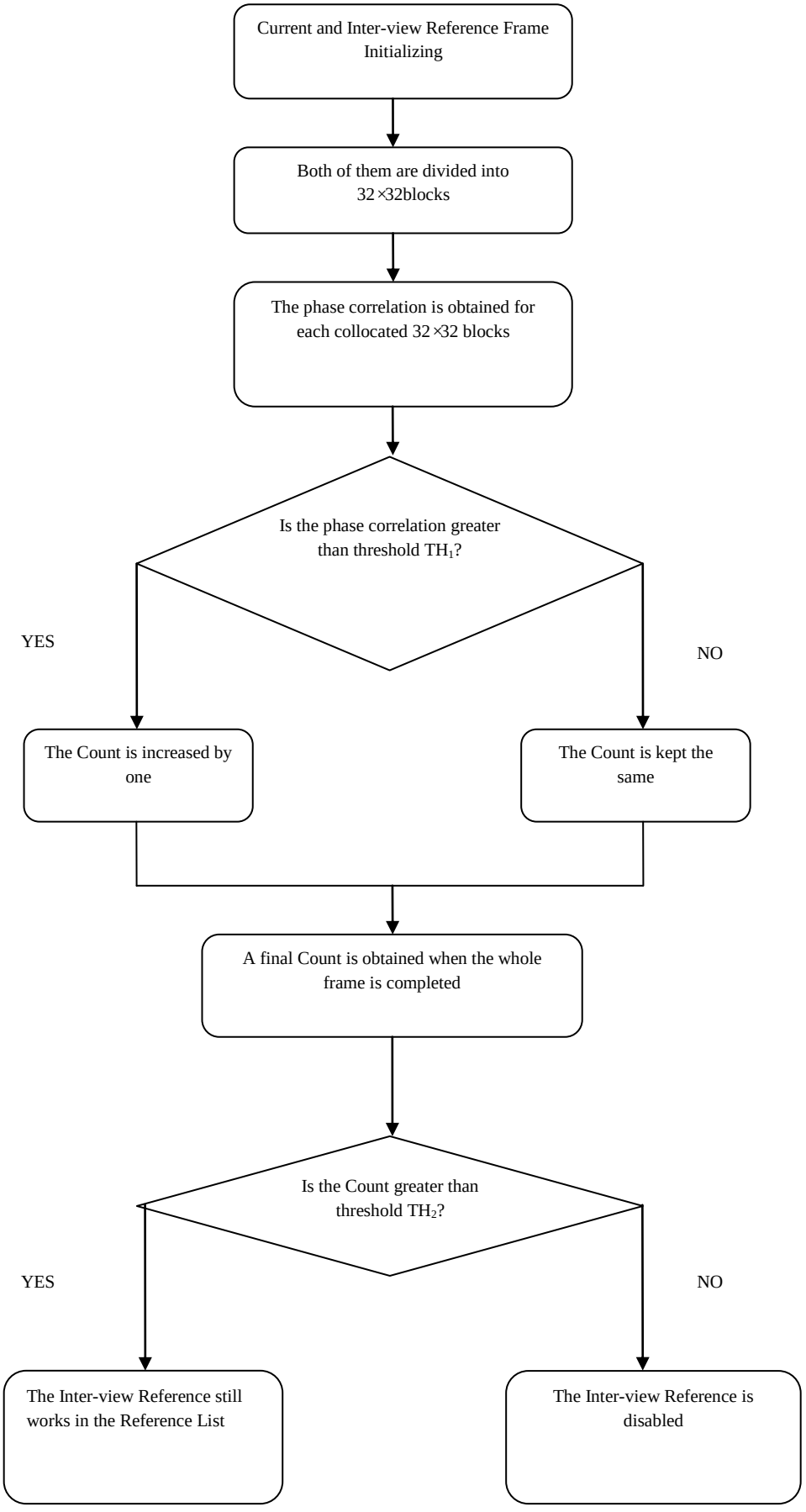


Figure 1 The flow chart of the inter-view reference frame skip decision

III. The experimental results and discussion

Our experiments are based on the JMVC version 8.5 of MVC reference software. “Ballroom”, “Exit” and “Race1” sequences are tested according to the common test configuration [9]. The test result of JMVC is shown in Table I. “DPSNR (dB)”, “DBR (%)”, “DT (%)” represents PSNR change, bit rate change in percentage and the entire coding time change in percentage.

Sequences	Picture Resolution	Proposed Method Compared with Standard JMVC		
		DPSNR (dB)	DBR (%)	DT (%)
Ballroom	640×480	-0.11	3.61	21.09
Exit	640×480	-0.02	0.77	19.41
Race1	640×480	-0.15	3.82	9.44
Average		-0.09	2.73	16.65

Table 16 Performance comparison between the proposed method and standard JMVC

It can be seen from Table I that the proposed algorithm reduce the encoding time with a negligible loss of coding efficiency in terms of quality and bit rate. The proposed method has reduced the encoding time by about 16.65% on average. The average DPSNR loss is 0.09 dB, and the increase of DBR is about 2.73% on average.

IV. Conclusions

This paper presents an inter-view reference frame skip decision algorithm for MVC. The dependence between the current frame and inter-view frame are decided based on the phase correlation of the sub-blocks. By definition, if the two frames with high correlation, then the inter-view reference frame will still work in the prediction reference list, otherwise, it will be disable in the list. Experimental results show that the proposed algorithm can reduce the computational complexity of MVC without complex search strategy or mode decision, and also maintain almost the same coding performance.

V. References

[1] A. Vetro, P. Pandit, H. Kimata and A. Smolic, "Joint Multiview Video Model (JMVM) 8.0," ISO/IEC JTC1/SC29/WG11 and ITU-T Q6/SG16, Doc. JVT-AA207, Apr. 2008.

[2] “Report on 3DAV Exploration”, ISO/IEC JTC/SC29/WG11, N5878, July 2003.

[3] “Description of Core Experiments in MVC”, ISO/IEC JTC1/SC29/WG11, MPEG2006/W7798, January 2006.

[4] L. Shen, Z. Liu, T.Yan, Z. Zhang, and P. An, “View-Adaptive Motion Estimation and Disparity Estimation for Low Complexity Multiview Video Coding,” IEEE Trans. Circuits Syst. Video Technol., vol. 20, no. 6, pp. 925-930, Jun. 2010.

[5] X. M. Li, D. B. Zhao, X. Y. Ji, Q. Wang, and W. Gao, “A fast inter frame prediction algorithm for multiview video coding,” in Proc. IEEE Int. Conf. Image Process. (ICIP), vol. 3, Sep. 2007, pp. 417–420.

[6] Pan Gao; Qiang Peng; Qionghua Wang; Xiangkai Liu; Chengde Zhang; , "Adaptive disparity and motion estimation for Multiview Video Coding," *Image and Signal Processing (CISP), 2011 4th International Congress on* , vol.1, no., pp.66-71, 15-17 Oct. 2011

[7] KUGLIN, C. D., AND HINES, D. C. 1975. The phase correlation image alignment method. In Proceedings of the IEEE 1975 International Conference on Cybernetics and Society (Sept.). IEEE, New York, pp. 163-165.

[8] M. Lukacs, “Predictive coding of multi-viewpoint image sets,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Tokyo, Apr. 1986, pp. 521–524.

[9] Y. Su, A. Vetro, and A. Smolic (July, 2006). Common Test Conditions for Multiview Video Coding ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, Doc. JVT-T207.

APPENDIX E

CD-ROM IN THE BACK POCKET

E.1 Scalable Video Coding (SVC) - Experimental Results.

E.2 Multiview Video Coding (MVC) - Experimental Results.

E.3 Ph.D Oral Defense - Presentation