3D wavelet based scalable video coding and its applications

Ni, Zefeng

2007

Ni, Z. (2007). 3D wavelet based scalable video coding and its applications. Master's thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/35736>

<https://doi.org/10.32657/10356/35736>

a530266

3D Wavelet Based Scalable Video Coding and Its Applications

Ni Zefeng

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in fulfilment of the requirement for the degree of
Master of Engineering

2007

Abstract

Scalable video coding is attractive due to the capability of reconstructing lower spatial resolution, or lower quality, or lower frame rate video signals from partial bit streams. It enables a simple solution in adaptation to different requirements from network, terminal and storage capabilities. In this thesis, a 3D wavelet based scalable video codec (SVC) is proposed through the integration of motion-compensated temporal filtering (MCTF) and JPEG2000. By exploring the unique features of MCTF and JPEG2000, the proposed simple codec not only achieves a comparable rate distortion (RD) performance but also has the properties of high scalability and error resilience. The main contributions of the developed codec lie in its simple architecture and the optimum bit allocation scheme based on the quality layer concept of JPEG2000.

For MCTF based video codecs, a fundamental question is how to allocate bits to each temporal band (T-band) generated by MCTF so that certain degree of constant quality can be achieved. In this thesis, a novel approach for constant quality aimed bit allocation is proposed for the applications of adaptive stored video streaming. The basic idea of the proposed bit allocation is to adjust the energy gains to compensate the different contributions from different types of T-bands, and more or less equally distribute the distortions among the T-bands at the same level. Experimental results show that the proposed bit allocation can greatly reduce the PSNR fluctuation with only slight degradation in average PSNR.

This thesis also investigates the potential usage for the developed SVC through designing two systems for streaming stored video over two types of networks, the best-effort Internet and QoS-enabled wireless network. For networks without QoS support, such as the present best-effort Internet, an end-to-end system is proposed. In particular, an end-to-end feedback control scheme is performed at the application layer to estimate the

available network bandwidth, which is then used to determine the video bit rate for the streaming session. For QoS-enabled networks, with the example of IEEE802.11e WLAN, a cross-layer design is proposed to optimize received video quality. Specially, at the application layer, the SVC stream is packetized with relative priority index (RPI) that marks the importance of video packets. At the network layer, available bandwidth is estimated and passed to the application layer to guide the truncation of video bit stream. In addition, adaptive QoS mapping is performed based on the RPI value of each packets. Both systems exploit the scalability property of SVC.

Acknowledgments

First and foremost the author would like to thank his supervisor, Assistant Professor *Cai Jianfei*, for his patient and altruistic assistance in the author's research and the preparation of this thesis. The author would also like to express his appreciation to his fellow graduate students in the Center for Multimedia and Network Technology for the valuable discussions with them.

Contents

Abstract	i
Acknowledgments	iii
List of Figures	vi
List of Tables	viii
LIST OF ABBREVIATIONS	ix
1 Introduction	1
1.1 Motivation & Objective	2
1.2 Organization of Thesis	4
2 Literature Review	5
2.1 Video Compression, Why and How?	5
2.2 Scalable Video Coding–Why?	9
2.3 Scalable Extensions to Hybrid Video Coding	11
2.3.1 Layered scalable coding without fine granularity	11
2.3.2 Layered coding with fine granularity–MPEG-4 FGS	13
2.3.3 Problem with existing scalable video coding	14
2.4 Wavelet Transform for Scalable Video Coding	16
2.4.1 Scalable Image Coding with Wavelet Transform	16
2.4.2 Inter-frame Wavelet Transform	20
2.4.3 3D Wavelet Decomposition for Scalable Video Coding	23
3 A Simple Scalable Video Codec Based on MCTF and JPEG2000	25
3.1 Structure of the Developed Scalable Video Codec	25
3.2 2D Decomposition and Spatial Coding	27

3.3	Rate control: optimum truncation across T-bands	28
3.3.1	Theoretical Analysis of MCTF	28
3.3.2	Bit Allocation Optimization Criteria	30
3.3.3	Preprocessing RD information	31
3.3.4	Finding optimum truncation points	32
3.4	Experimental Results	32
3.5	Summary	36
4	Constant Quality Aimed Bit Allocation for 3D Wavelet Based SVC	38
4.1	Problem Statement	40
4.2	Constant Quality Aimed Bit Allocation	43
4.3	Experimental Results	48
4.4	Summary	52
5	Application of 3D Wavelet Based SVC for Video Streaming	53
5.1	An End-to-end Video Streaming System for Best-effort Internet	53
5.1.1	Background	54
5.1.2	End-to-end System Design	57
5.1.3	Experimental Results	59
5.2	Video streaming Over QoS Enabled Network, IEEE 802.11e WLAN	62
5.2.1	Review on IEEE 802.11e WLANs	63
5.2.2	Optimized Cross Layer Design	63
5.2.3	Application Layer: Macro Rate Control and Priority Assignment	67
5.2.4	Network Layer: Bandwidth Estimation and Adaptive QoS Mapping	69
5.2.5	Experimental Results	72
5.3	Summary	76
6	Conclusions and Future Work	77
6.1	Future Work	79
	References	81
	Related Publications	88

List of Figures

2.1	A typical encoding structure for hybrid video coding	7
2.2	Frame types and their relationships with regard to prediction	8
2.3	System configuration for streaming video application	9
2.4	Illustration of video coding performance	10
2.5	Some layered video coding structures	13
2.6	Hybrid temporal and SNR scalability with FGS	14
2.7	Illustration of 2D discrete wavelet transform (DWT)	17
2.8	Code-stream partitioning in EBCOT	19
2.9	Multiple levels of temporal wavelet decomposition	20
2.10	Two levels of MCTF with 5/3 wavelet	22
2.11	3D wavelet cube for 3D wavelet decomposition	24
3.1	The encoding structure	26
3.2	Probability density of two T-bands of the Y component	27
3.3	Convex hull interpretation	31
3.4	Finding the optimum truncation points	33
3.5	RD performance in terms of PSNR	34
3.6	Sample reconstructed frames	35
3.7	ROI coding (left: 23kbps without ROI, right: 23kbps with ROI coding)	37
4.1	Distortion relationship between frame L_k and frame B_k	40
4.2	PSNR estimated with conventional MCTF distortion analysis.	41
4.3	PSNR with MC-EZBC for Stefan_CIF at 1.2Mbps with two levels of MCTF.	42
4.4	PSNR with the developed JPEG2000 based codec for Stefan_CIF at 1Mbps with two levels of MCTF.	42

4.5 Typical MSE for each T-band with the JPEG2000-like optimum bit allocation 43

4.6 Smoothing MSE of the T-bands at the same level for Stefan_CIF at 1Mbps . . 45

4.7 Estimating PSNR with the proposed model (Stefan_CIF at 800kbps) . . 47

4.8 Estimating average PSNR with different models 47

4.9 Constant quality aimed bit allocation for Stefan_CIF at 1Mbps 49

4.10 Experimental results for Foreman_CIF at 400kbps 50

4.11 Experimental results for Table_tennis_CIF at 600kbps 51

5.1 The System architecture for the end-to-end streaming system 56

5.2 A peer-to-peer network 60

5.3 Performance of the end-to-end streaming system 61

5.4 The four access categories in an EDCA node 64

5.5 Cross layer design for video streaming over IEEE 802.11e WLAN 66

5.6 An example of the loss impact results 69

5.7 An example of the RPI value for each packet 69

5.8 WLAN experiment setup and PSNR for different streaming settings . . . 73

5.9 Packet loss ratio for different streaming settings on IEEE 802.11e WLAN 75

5.10 Illustration of macro rate control in the cross-layer design 75

List of Tables

4.1 Comparison of variance and average of the PSNR between the proposed constant quality aimed bit allocation scheme and the JPEG2000-like bit allocation scheme. 51

List of Abbreviations

Abbreviations	Details
AC	Access Category
AIMD	Additive Increase and Multiplicative Decrease
AIR	Additive Increase Rate
CBR	Constant Bit Rate
CIF	Common Intermediate Format (352 × 288)
DCF	Distributed Coordination Function
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
EBCOT	Embedded Block Coding with Optimal Truncation
EDCA	Enhanced Distributed Channel Access
EZW	Embedded Zerotree Wavelet
FEC	Forward Error Correction
FGS	Fine Granularity Scalability
GOP	Group Of Pictures
HCF	Hybrid Coordination Function
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IP	Internet Protocol
IPv4	Internet Protocol Version 4
ISO	International Standards Organization
ITU	International Telecommunication Union
ITU-T	ITU Telecommunication Standardization Sector
JPEG	Joint Photographic Experts Group
LAN	Local Area Network
MAC	Medium Access Control
MB	Macroblock
MC	Motion Compensation
MC-EZBC	Motion Compensated - Embedded Zero Block Context
MCTF	Motion Compensated Temporal Filtering
MDF	Multiplicative Decrease Factor
ME	Motion Estimation

Abbreviations	Details
MIMD	Multiplicative Increase and Multiplicative Decrease
MPEG	Moving Picture Experts Group
MSE	Mean Square Error
MTU	Maximum Transit Unit
MV	Motion Vector
PSNR	Peak Signal-to-Noise Ratio
QCIF	Quarter Common Intermediate Format (176)
QoS	Quality of Service
R-D	Rate-Distortion
RLI	Relative Loss Impact
ROI	Region of Interest
RPI	Relative Priority Index
RTCP	Real-time Control Protocol
RTP	Real-time Transport Protocol
RTT	Round Trip Time
SNR	Signal-to-Noise Ratio
SPIHT	Set Partition In Hierarchical Trees
SVC	Scalable Video Codec
T-band	Temporal Subband
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
UEP	Unequal Error Protection
ULP	Unequal Loss Protection
VBR	Variable Bit Rate
VLC	Variable Length Coding
WAN	Wide Area Network
WLAN	Wireless Local Area Network

Chapter 1

Introduction

Digital video compression techniques have played an important role in the world of telecommunication and multimedia systems where bandwidth is still a valuable commodity. Hence, video coding techniques are of prime importance for reducing the amount of information needed for a picture sequence without losing much of its quality, judged by the human viewer [1]. Since the early 1990s, we have experienced a great advancement for digital video compression. Many video coding methods has been standardized internationally, such as MPEG-1, MPEG-2, H.263 etc. However, most of standardized video coding methods do not deal with the diversified requirements for video applications. In the future, video will often be transmitted over variable channels with variable bandwidth, both in wireless and wired networks. They need to be stored on media with different capacity, from flash memory disks to high-capacity DVD. They have to be decoded and displayed on a variety of devices, from mobile phones with limited processing power and small screen size to high-resolution projection system connected to a powerful dedicated video processor. One simple solution is to encode the same video many times for different potential applications. However, since there are usually so many potential applications, to deal with each one of them individually is very difficult. In addition, this will require a lot of processing at server/network and cause high-complexity for decoding. Another method is to use scalable video coding.

CHAPTER 1. INTRODUCTION

Scalable video coding schemes encode video signal once at highest resolution with highest quality, but enable decoding from partial streams depending on the specific requirement by a certain application. This enables a simple and flexible solution for transmission over heterogeneous networks and provides adaptability for bandwidth variations and error conditions. Applications based on these schemes are possible with minimal processing at server/network and low-complexity decoding. In addition, it allows simple adaptation for a variety of storage devices and terminals [2]. Typically, scalable coding schemes provide different scalabilities with fine granularity at bitstream level, including spatial (i.e. different display resolution), temporal (i.e. frame rate) and quality-level resolutions. The latter is often referred to as signal-to-noise (SNR) scalability.

The main contribution of this thesis includes the development of a simple scalable video codec (SVC) based on 3D wavelet transform and a novel constant quality aimed bit allocation algorithm. In addition, the potential usage for the developed SVC is also investigated with two streaming systems for streaming stored video over two kinds of networks, best-effort Internet and QoS-enabled wireless network.

1.1 Motivation & Objective

Numerous scalable coding schemes have been proposed to be integrated into today's video coding standard. However, a wide acceptance in the market of prospective applications has never occurred. The reason is the lack of coding efficiency in combining scalable coding with conventional hybrid motion-compensated prediction and block transform encoding. This is mainly caused by the recursive structure of the prediction loop, which causes a drift problem whenever incomplete information is decoded. Hence, research for more efficient scalable coding techniques is a demanding topic in digital video compression. Provisions are possible which minimize the effect of drift by modifying the structure

CHAPTER 1. INTRODUCTION

of the prediction loop. For example, the recent breakthroughs in motion-compensated temporal filtering (MCTF) [3, 4, 5] based on wavelet transform, which entirely abandons the recursion in encoding and decoding, have finally enabled the implementation of highly scalable video codecs [1]. At the same time, 2D wavelet transform based image coding schemes [6] have been well established in the last several years. As a consequence, wavelet based video coding schemes can provide flexible spatial, temporal and SNR scalability with fine granularity over a large range of bit rates while the compression performance does not have to suffer much. This feature of wavelet transform is demonstrated in this thesis by the development of such a scalable coding scheme based on the integration of MCTF and JPEG2000.

Scalable video coding scheme, include the proposed MCTF and JPEG2000 based codec, encodes the video at the highest quality, i.e. at a very high bit rate. Given a specific bit rate, only part of the bit stream will be transmitted or saved for decoding. Which part of the stream shall be decoded will directly affect the quality of the decoded video, i.e. some kind of rate optimization has to been applied. For the developed MCTF and JPEG2000 based codec, a optimized rate allocation based on theoretical model analysis is proposed. In additional, a new rate allocation method (here after called constant quality aimed bit allocation) is also proposed to reduce the quality fluctuation in the reconstructed video frames, which often exists in theoretically optimum rate allocation method. There could be many potential application for the 3D wavelet based SVC. One popular example is its ability to enable easy adaptability for bandwidth variations and error conditions when transmitting video streaming over heterogeneous networks. This issue will also be discussed in this thesis. Two examples of streaming system are presented with the developed SVC.

1.2 Organization of Thesis

The organization of the thesis is as follows. Chapter 2 outlines the principles for digital video compression. It discusses various conventional hybrid video coding techniques as well as some of their scalability extensions. 2D Wavelet transform and its application on scalable still image coding will also be introduced with focus on the JPEG2000 standard. Typical scalable video coding schemes based on 3D wavelet transforms are also reviewed. Chapter 3 documents the development of the proposed scalable video codec based on MCTF and JPEG2000. Chapter 4 describes a simple and effective algorithm on constant quality aimed bit allocation for the proposed scalable video codec. Chapter 5 presents two streaming system, for the best-effort Internet and QoS-enabled wireless LAN respectively, to illustrate the potential usage for the developed 3D wavelet based scalable codec. Chapter 6 concludes this thesis and outlines some possible future work.

Chapter 2

Literature Review

Since raw video signal contains a large amount of information, compression is essential for efficient storage and transmission. In this chapter, some typical video compression schemes will be reviewed, including traditional non-scalable hybrid video encoding like MPEG-2, and their scalability extensions. In addition, some relatively new coding methods based on wavelet transform will be surveyed.

2.1 Video Compression, Why and How?

A moving scene (i.e. a video stream) is a sequence of individual frames. Every frame is a still image which when shown together, one after another, become a moving video sequence. Raw digital video places a heavy burden for storage or transmission. To appreciate this, one only needs to calculate the amount of storage space required by raw video. Take the picture represented with CIF (Common Intermediate Format) [1] for example:

The picture resolution is 352×288 . All luminance information is retained while chrominance information is sub-sampled 2:1 in both horizontal and vertical directions. This image format is also known as 4:2:0 format. Each sample of the YUV components is represented with 8 bits. The frame rate is 30 frames per second. The resulting bit rate for this CIF sequence is then:

CHAPTER 2. LITERATURE REVIEW

- Luminance (Y): $352 \times 288 \times 30 \times 8 = 24,330,240$ bps (bits per second)
- Chrominance (U): $352 \times 288 \times 30 \times 8/4 = 6,082,560$ bps
- Chrominance (V): $352 \times 288 \times 30 \times 8/4 = 6,082,560$ bps

The total bit rate for this raw video is $36,495,360 \approx 34.8$ Mbps, which is hardly achieved by today's wide area network. One hour of such video would require about 122 Gigabytes of storage. Clearly, for practical storage or transmission, some way of compression is necessary to reduce the amount of bits required to represent a video sequence. Luckily, for video signals, there is a strong correlation between successive picture frames and within the picture elements themselves. In other words, the raw digital video contains a lot of redundant information, both within each frame and between frames. Theoretically de-correlation of the video signals can compress bandwidth without significantly reducing image quality. Moreover, the fact that the human beings are insensitive to loss of certain spatial-temporal visual information can be exploited for further reduction in bit rate by applying some subjectively lossy compression techniques [1]. Since the early 1990s, many video coding methods have been proposed. All of them, including the standard video codecs, such as H.261 [7], MPEG-1 [8] and MPEG-4 [9] etc., exploit both the spatial correlation and the temporal correlation. They are all based on three fundamental redundancy reduction principles:

- (i) *Spatial redundancy reduction*: to reduce spatial redundancy among the pixels within a picture (i.e. the similarity of pixels within a frame), by employing some data compression techniques, such as transform coding which maps the pixels into a transform domain prior to data reduction. The reason is that transformation causes the significant part of the image energy to be concentrated in the low frequency region and hence into a few transform coefficients, which can then be quantized by discarding the unimportant coefficients having little energy without significantly

CHAPTER 2. LITERATURE REVIEW

Figure 2.1: A typical encoding structure for hybrid video coding

affecting quality of the reconstructed image. Note that this is a lossy process and cannot be reversed without certain level of quality degrading.

- (ii) *Temporal redundancy reduction*: to remove the similarities between successive frames, by coding their differences. This is called inter-frame coding. Usually image changes due to motion can be significantly reduced if the motion of objects can be estimated and the difference is taken on the motion compensated image.
- (iii) *Entropy coding*: to reduce redundancy between the compressed data symbols, using variable length coding techniques. The two most common known methods are Huffman coding or arithmetic coding, which are based on probability statistics.

Most international standardized video codecs, either the ITU-T H.261 [7], H.263 [10], or MPEG-1 [8], MPEG-2 [11] and MPEG-4 [9] from ISO/IEC, are often called hybrid video coding which combines together two stages: first, motion estimation (ME) and motion compensation (MC) predict each frame from the neighboring frames, then the

Figure 2.2: Frame types and their relationships with regard to prediction

prediction error is further encoded to reduce the correlation in the spatial domain. Figure 2.1 shows a general overview of the MPEG-4 encoder, a typical hybrid video encoder, which uses discrete cosine transform (DCT) and quantization of DCT coefficients to remove the spatial redundancy within a frame. Temporal redundancy is reduced by the use of inter-frame coding and motion compensation. Three different types of frame are specified: I-frame, or intra coded frame, which is coded without reference to any other pictures; P-frame, or predictive coded frame, which is coded using motion compensation from a previous I-frame or P-frame; B-frame, or bidirectional predictive coded frame, which is coded using motion compensation from a previous and a future I-frame or P-frame. All P-frames and B-frames rely directly or indirectly on I-frame to be decoded successfully (see Figure 2.2). Even the latest international video coding standard, ITU-T H.264 (also known as the ISO/IEC MPEG-4 Part 10 advanced video coding, i.e. H.264/AVC) [12], which has a much higher coding efficiency than all its predecessors, follows a similar coding structure. The main difference is the way of motion estimation (use of multiple reference frames, spatial transform (an integer transform rather than DCT) and the use of arithmetic coding instead of Huffman coding.

Figure 2.3: System configuration for streaming video application

2.2 Scalable Video Coding—Why?

Traditionally the objective of video coding is to optimize video quality at a given bit rate. Its performance is often called rate-distortion (RD) performance, i.e. the distortion (quality)¹ of the reconstructed video at different bit rates. For hybrid video coding methods, its RD performance has improved dramatically with the introduction of MPEG-4 and H.264. However, it is non-scalable. The quantization step size for DCT (or some other integer transform) coefficients will directly determine the quality of the reconstructed video. When a video is encoded, its rate and distortion has been fixed after the encoding process. If a lower bit-rate (or lower quality) stream for the same video is required due to bandwidth or storage constraint, the only way is to decode the video stream and re-encode the reconstructed video at a lower rate, a process which is called transcoding. However, transcoding not only introduces additional computation but also causes a drop in RD performance (i.e. transcoding at a lower rate has a lower RD performance than encoding the raw video directly at the lower rate) due to the nature of the lossy compression.

Why for the same video, different bit rates are often required? Consider a typical system for network video application (Figure 2.3), the network channel capacity varies over a wide range depending on the type of connection and the network traffic at any

¹Typically the distortion is measured by Mean Square Error (MSE) due to its easiness of calculation and the tractability of linear optimization problems involving squared error metrics. The lower the MSE is, the lower the distortion will be, i.e the better the quality is. Very often for image/video compression, the MSE is commonly quoted in terms of the equivalent reciprocal measure, Peak Signal to Noise Ratio (PSNR), which is expressed in dB (decibels). Good reconstructed frames typically have PSNR values of 30dB or more.

Figure 2.4: Illustration of video coding performance [13]

given time. In addition, video decoder at the client terminal may not be able to decode all the bits received from the channel fast enough for reconstruction due to its own hardware/software limitation. Therefore, it is required that the video server is capable to stream the same video at various bit rate. Transcoding is one way to achieve this at a very high computational cost. Another way is to use scalable video coding, i.e. encode the video at highest quality and rate while allowing the bitstream to be partially decodable at any bit rate within a bit rate range for reconstruction with an optimized quality at that bit rate. Figure 2.4 illustrates this point.

The RD curve in Figure 2.4 indicates the upper bound in quality for any coding technique at any given bit rate. The two staircase curves indicate the performance of an optimal non-scalable coding technique. Once a given bit rate is chosen—either low or high—the non-scalable coding technique tries to achieve the optimal quality by having the upper corner of the staircase curve very close to the RD curve. If channel bit rate is lower than the video bit rate, a so-called “digital cutoff” phenomenon happens and the received video quality becomes very poor. On the other hand, if the channel bit rate is

CHAPTER 2. LITERATURE REVIEW

higher than the video bit rate, the received video quality does not become any better. Therefore, scalable video coding supporting variable bit rates will be very useful in this case in terms of network utilization and quality of service at the client side.

Note that Figure 2.4 only illustrates the case for SNR scalability, where frame rate and spatial resolution are fixed. There are two other popular scalabilities for video coding, namely, temporal and spatial scalability.

- Temporal scalability is a technique to code a video sequence into several layers at the same spatial resolution but with different frame rates. Temporal scalability is useful because video clients could have different requirements on frame rate since frame rate could directly affect the computational load for decoding. Sometimes client would also like to drop frame rate in trade for higher quality in individual frames, which is beneficial for low-motion video sequence.
- Spatial scalability is a technique to code a video sequence into several layers at the same frame rate but different spatial resolutions. For example, a 4CIF sequence could be encoded with three layers with resolution 4CIF, CIF and QCIF respectively. Client might have different resolution needs because of its own displaying capability, e.g. mobile devices such as hand-phone typically have very limited screen resolution. Very often, the selection of resolution will also be determined by human viewers themselves (i.e. different people could have different preference for the same video).

2.3 Scalable Extensions to Hybrid Video Coding

2.3.1 Layered scalable coding without fine granularity

Due to the above scenario, scalable video coding has always been an interesting topic. In MPEG-2 and MPEG-4, layered scalability techniques, including SNR scalability, temporal scalability and spatial scalability, have been included. In such a technique, video

CHAPTER 2. LITERATURE REVIEW

is coded into a base layer and an enhancement layer. The enhancement layer bitstream is similar to the base layer bitstream in the sense that it has to be either completely received and decoded or it does not enhance the video quality. It changes the non-scalable single staircase curve to a curve with two stairs (see Figure 2.4). The base-layer bit rate determines the position of first stair and the enhancement layer bit rate determines the second stair. Figure 2.5(b) shows such a layered encoding structure, where the enhancement layer simply re-encode the quantization distortion of the base layer transform coefficients with a finer quantization step size. For temporal scalability, typically the enhancement layer provides the missing frames to form a video with a higher frame rate. Figure 2.5(a)) shows such a structure where only P-type motion compensation (prediction) is applied in base layer. For the enhancement layer prediction, either P-type or B-type from the base layer or P-type from the enhancement layer can be used. For spatial scalability, usually the reconstructed based-layer picture is up-sampled to form the prediction for the high-resolution picture in the enhancement layer. Figure 2.5(c) shows such an encoding structure with two layers.

The advantage with layered coding is its simplicity. It intuitively extends hybrid video coding like MPEG-4 with scalability by introducing enhancement layer(s). However, since each layer has fixed bit-rate, the kind of scalability provided by layered coding methods is very limited because the enhancement layer is either entirely transmitted/received/decoded or it does not provide any enhancement at all. For example, for SNR scalability, which only provides support for several bit rate, a potential application could require a large variety of bit rates which are unknown during the encoding process which makes it difficult to select the bit rates for the base and enhancement layer. In one word, what a typically application (such as network streaming) desired for video coding is a continuous curve that parallels the RD curve with a single stream (see Figure 2.4). This

CHAPTER 2. LITERATURE REVIEW

Figure 2.5: Some layered video coding structures

is the objective of the fine granularity scalability (FGS) video coding technique in the Amendment of MPEG-4 [14, 13].

2.3.2 Layered coding with fine granularity–MPEG-4 FGS

MPEG-4 FGS aims to provide a continuous scalability curve illustrated in Figure 2.4. The main difference between FGS and the other layered method discussed above is that, although the FGS coding technique also codes a video sequence into two layers (the base layer coding is unchanged), its enhancement layer bitstream can be truncated into any number of bits within each frame to provide partial enhancement proportional to the number of bits decoded for each frame [14]. Because the enhancement layer in FGS encodes the difference between the original picture and reconstructed picture from the base layer using *bit plane coding* of DCT coefficients, the decoder of FGS could then reconstruct an enhancement video from the base layer and a truncated enhancement layer bitstream after encoding is completed. Bit plane coding could be considered as a kind of quantization with *successive approximation* (i.e. the quantization step for the DCT coefficients is gradually refined with the decoding of each bit plane). The more

CHAPTER 2. LITERATURE REVIEW

Figure 2.6: Hybrid temporal and SNR scalability with FGS

number of bits decoded for each picture, the more bit planes will be reconstructed (i.e. reconstructed DCT coefficients will more precise), resulting in better video quality.

Because of the fine granularity in bit plane coding, it has also been used to provide temporal and spatial scalability, such as the hybrid temporal-SNR FGS in [15] and the hybrid spatial-SNR FGS in [16]. Figure 2.6 shows a sample configuration which supports two enhancement layers with fine granularity in hybrid SNR and temporal scalability.

2.3.3 Problem with existing scalable video coding

Layered coding like MPEG-4 FGS could provide fine scalability, which solves the problem of limited scalability in other layered coding. However, it has not yet widely used in applications because of the following drawbacks:

- Difficult to achieve hybrid scalability. It might appear straight-forward to combine individual scalability to form hybrid scalability by introducing multiple enhancement layers as Figure 2.6. However, when there are several temporal scalability (e.g. 30, 15 and 5 frames per second) and several spatial scalability (e.g. 4CIF, CIF and QCIF), how to arrange the order for each enhancement layer is not easy. The bitstream structure might become very complicated. Which is why none of the video coding standards existing so far restricts scalability at the bitstream level to a predefined number of layers, which must be known at the time of encoding [2].

CHAPTER 2. LITERATURE REVIEW

- Trade off between RD performance and drift effect for SNR scalability: In ideal case, the base layer and any composition from layers should be self-contained, which means that the prediction should not use any decoded information from higher layers. Consider MPEG-4 FGS or the layered structure in Figure 2.5(b), the reconstructed base layer frame is used for prediction since the encoder and decoder must use same reference frame due to recursive frame prediction loop in hybrid video coding concept. Otherwise, different estimates would be used at the encoder and decoder sides and a drift effect would then occur. This does not penalize the operation of the coder at the base layer, which performs like a single layer coder at the same rate. However, the RD performance toward higher bit rates will be worse than what it could be in a single layer coder because the prediction from base layer information will always be worse than that it could be if all or part of enhancement layer information was allowed in the prediction. Alternatively, if full enhancement information is blindly used for prediction, which is actually a mode of SNR scalability in MPEG-2 [11], the reconstructed quality of the highest enhancement layer could approach the performance of a single layer coder while the reconstructed quality of the base layer and all intermediate bit rate (for layers with fine granularity) would eventually suffer dramatically due to the drift. This dilemma to penalize either enhancement or base layer is inevitable. Many attempts have been done to optimize the trade off for different sequences under the constraints of expected rate targets, such as the progressive FGS in [17] and [18], Trellis-based predictive Leak in [19]. Nevertheless, their RD performance is still worse compared with single layer coders due to the recursive structure and drift problem. Further more, these methods often result in systems of considerable complexity both at the encoder and decoder side [2]. To overcome this dilemma, it would be then desirable to discard the temporal prediction recursion. One approach is to use an open loop temporal wavelet transform, which is to be discussed in the following section.

2.4 Wavelet Transform for Scalable Video Coding

Discrete wavelet transform, under the general name of sub-band coding, was first introduced by Crochiere et al. in 1976 [20], and has since proved to be a simple and powerful technique for speech and image compression. The basic principle is the partitioning of the signal spectrum into several frequency bands and the separate coding and transmission for each band. This is particular suited for image coding. Before it is used for video coding, it has already been extensively used for image coding, which even results in international standard like the coding of still images in JPEG2000 [6].

2.4.1 Scalable Image Coding with Wavelet Transform

Natural images tend to have non-uniform frequency spectrum, with most of the energy being concentrated in the lower frequency band. Furthermore, according to human vision system, noise visibility tends to fall off at both high and low frequencies and this enables the adjustment of compression distortion based on perceptual criteria. In addition, wavelet transform processes image in their entirety, unlike the DCT transform using artificial blocks, therefore there is no block artifacts in the coded picture, as occurs in the DCT based image coder like JPEG.

In wavelet coding, the band splitting is done by passing the image data through a bank of bandpass filters, which is often arranged into octave-bands in order to adapt the frequency response of the decomposed pictures to the characteristics of the human visual system [1], which is often referred to as “dyadic” wavelet decomposition. This process is often called wavelet decomposition, as illustrated in Figure 2.7. $*_h$ and $*_v$ denote horizontal and vertical convolution along rows and columns of the image, respectively. h_0 and h_1 are the low pass and high pass filters, respectively. Therefore $*_h h_0$ means filtering with h_0 along the rows of the image. Similarly, $\downarrow_h$ and $\downarrow_v$ denote horizontal and vertical down-sampling (decimation). “ HL ” is used to identify the sub-band with

CHAPTER 2. LITERATURE REVIEW

Figure 2.7: Illustration of 2D discrete wavelet transform (DWT)

horizontal details (i.e. high frequency content in the vertical direction and low frequency content in the horizontal direction). Similarly, the sub-band LL , LH and HH band are defined. Obviously the LL band could be considered as the same image with lower resolution, which is why wavelet transform has a natural spatial scalability (for image coding, this scalability is often called resolution scalability). Typically, further wavelet decomposition is applied for the low pass band recursively, which results in a multi-band, tree-structured wavelet decomposition (filter bank). For wavelet decomposition with D levels, it yields $3D + 1$ sub-bands, denoted as $LH_0, LH_0, HH_0, LH_1, \dots, HH_{D-1}, LL_{D-1}$. This structure is called multi-resolution in [21] because once the sub-bands are encoded separately it will provide a mechanism whereby a compressed bit stream may be partially decompressed to obtain successively higher resolution versions of the original image, i.e. high spatial scalability.

For efficient compression of the image as well as for a wide range of scalability, the wavelet coefficients are often bit plane coded, which often has less coding efficiency than

CHAPTER 2. LITERATURE REVIEW

conventional quantization and entropy coding. However, in [22] Shapiro introduces an embedded zero tree coding (EZW) which produces state-of-art RD performance while maintaining the SNR scalability property of bit plane coding. The bit plane coding of EZW allows data dependent departures from conventional raster scan of transform coefficients. These departures are in the form of “zero” tree, which exploit the similarities of the sub-bands of the same orientation for the coding of large number of zeros using very few compressed bits. Due to its great RD performance, modest complexity and scalable stream structure, EZW captured the interest of the compression research community and a tremendous research effort was ignited by [22]. In [23], Shapiro’s zero-trees were generalized and set partitioning techniques (named set partition in hierarchical trees–SPIHT) were introduced to effectively code these generalized trees of zeros, which produces superior result compared with EZW at even lower levels of complexity. SPIHT then quickly became the yardstick by which all wavelet-based image compression methods are measured.

Though zero-tree based method has amazing RD performance and SNR scalability, it has one draw back – the loss of spatial (resolution) scalability in its bit stream structure since all sub-bands are encoded together rather than separately. In order to encode the wavelet coefficients in a bit stream with both spatial and SNR scalability, Taubman introduces the the concept of *Embedded Block Coding with Optimal Truncation (EBCOT)* [24], which is adopted by the JPEG2000 [6] standard. Its RD performance is comparable to SPIHT scheme with extra functionalities including both SNR and resolution scalability, error resilience and region of interest (ROI) coding. In EBCOT, each sub-band is partitioned into relative small blocks (say 32×32 samples each, called “code-blocks”), each of which is coded independently using bit-plane coding. Therefore the bit stream is both SNR and spatial scalable. Furthermore, the independent coding of code-block also makes the final bit-stream error resilient.

Figure 2.8: Code-stream partitioning in EBCOT

Quality scalability in EBCOT is achieved through the use of “quality layers”. Each code-block bit stream is embedded encoded with a large number of truncation points (i.e. highly scalable due to bit plane coding), each of which will contribute a certain distortion level and bit length. For a quality layer with a certain rate R , the optimum truncation method defined in EBCOT optimum select truncation points from all the code-blocks and pack them together as illustrated in Figure 2.8. Five layers and six code-blocks are shown in Figure 2.8. Each rectangle represents the chunks of a code-block stream (at the truncation points) which have been included in the relative layer. The data coded for a block is represented as the chunks lying on the corresponding vertical arrow, while the compressed bitstream is transmitted/stored in a per-layer fashion (i.e. lower layers first). Typically a final encoded bit stream with JPEG2000 is composed of a collection of such “quality layers”. This optimum truncation also helps to improve the RD performance of the codec.

Figure 2.9: Multiple levels of temporal wavelet decomposition

2.4.2 Inter-frame Wavelet Transform

A video sequence is basically a collection of still related images. 2D wavelet transform is useful for the de-correlation of 2D spatial image and further compression. To compress a video with wavelet transform, one more dimension of wavelet transform is required to capture the correlation between neighboring frames, i.e. temporal redundancy, that is temporal filtering or wavelet transform along the temporal axis. Figure 2.9 shows such a wavelet decomposition tree with Harr filter. This type of schemes without motion compensation have in fact been proposed more than 15 years ago, e.g. [25]. In the case of a non-orthonormal transform, this can be interpreted as decomposition of a frame pair (A,B) into one average (low pass) and one difference (high-pass) frame described as

$$L(m, n) = \frac{1}{2}[A(m, n) + B(m, n)]; \quad (\text{Eq. 2.1})$$

$$H(m, n) = A(m, n) - B(m, n). \quad (\text{Eq. 2.2})$$

If pairs of low-pass frames are then again combined, subsequence levels of a wavelet tree could be established as in Figure 2.9. Most early works on 3D wavelet based video coding usually apply simple temporal filtering (typically 2-tap Harr filter). For example,

CHAPTER 2. LITERATURE REVIEW

the work in [26] did not use motion compensation (MC), while the work in [27] introduced the concept of “connected” (area along the motion trajectory) and “unconnected” (cover and uncover areas that are not members of motion trajectory) to enable the application of MC in the temporal filtering, which is further improved by [28].

Application of MC is the key for high compression performance in video coding. It is often understood to be implicitly coupled with frame prediction schemes as in hybrid video coding. However, there is indeed no justification for this restriction since MC can rather be interpreted as a method to align a filtering operation along temporal axis with a motion trajectory [2]. In the case of hybrid video coding with MC prediction, the filters are in principle linear predictive coding analysis and synthesis filters, while in the cases of wavelet transform, transform basis functions extended over the temporal axis are subject to MC alignment, i.e. linear filtering operation is done along a motion trajectory. This process is known as *Motion compensated temporal filtering (MCTF)*. When MC is used for wavelet transform, it reduces the texture details in high pass band, the same behavior as a low prediction error when MC is used for hybrid coding. In addition, the low pass frame will then contain all relevant image information; it appears similar to an original frame, but indeed is an average over many frames depending on the level of temporal filtering, which is why MCTF is extremely useful for the purpose of temporal scalability when temporal low bands (e.g. L, LL, etc.) could be considered as the same original video but with lower frame rate.

In [3], [4] and [5], the lifting scheme was proposed for the implementation of MCTF, which allows efficient implementation of temporal filtering with high order. Figure 2.10 shows two levels temporal decomposition with MCTF using the popular 5/3 wavelet. In particular, MCTF is implemented by the 2-step lifting scheme, which consists of the prediction and the update steps. The former takes the original input frames to generate

Figure 2.10: Two levels of MCTF with 5/3 wavelet

the high-pass bands H_k ; the latter uses the available high-pass bands and even frames to generate the low-pass bands L_k . The lifting steps are as follows:

$$H_k[m, n] = F_{2k+1}[m, n] - 0.5(W_{2k,2k+1}(F_{2k})[m, n] + W_{2k+2,2k+1}(F_{2k+2}[m, n])), \quad (\text{Eq. 2.3})$$

and

$$L_k[m, n] = F_{2k}[m, n] + 0.25(W_{2k-1,2k}(H_{k-1})[m, n] + W_{2k+1,2k}(H_k[m, n])), \quad (\text{Eq. 2.4})$$

where W_{1-2} denotes the motion compensated mapping of the first frame onto the coordinate system of the second, so that $W_{1-2}[m, n] \approx F_2[m, n]$. In the prediction step, bi-direction ME is performed. Note that any kinds of ME/MC methods could be used here including the conventional block-based motion compensation used in hybrid video coding like MPEG-4 and H.264. In the update step, pixels will be updated by the high-pass coefficients that are predicted with these pixels with the corresponding weighting factor. It guarantees the match in the predicting and updating steps. Note that the presented structure is the same as an open-loop version of a hybrid video codec if the update step is removed.

Typically multiple levels of MCTF is applied for a larger set of frame rates. This, as a by-product, enables the exploration of multiple-frame redundancies which is hardly

CHAPTER 2. LITERATURE REVIEW

achievable by the conventional frame-to-frame in MPEG-4 or the multi-frame prediction in H.264 [2], this is because the low-pass frames generated by MCTF highlight those information parts of the video which are consistent over a large number of frames. In addition, the open-loop structure (i.e. no recursive prediction) of MCTF could avoid the drift problem (see section 2.3.3) typically associated with hybrid coders. Therefore MCTF is widely used for wavelet based scalable video coding.

2.4.3 3D Wavelet Decomposition for Scalable Video Coding

As discussed, coding technique based on 2D wavelet transform such as JPEG2000 could provide both SNR and spatial scalability for still image coding while MCTF introduces an efficient way for temporal scalability. Combined together, they would generate a 3D (spatial temporal) wavelet transform tree which could provide SNR, spatial and temporal scalability. One easy way is to apply the 2D wavelet decomposition to temporal band (T-band) after MCTF (a process often called $T+2D$). Typically such a 3D wavelet will generate many spatial-temporal sub-band as illustrated in Figure 2.11. Clearly, if each sub-band is embedded coded for SNR scalability, combined with the inherent temporal and spatial scalability in 3D wavelet, a highly scalable video coding technique will then be possible.

CHAPTER 2. LITERATURE REVIEW

Figure 2.11: 3D wavelet cube for 3D wavelet decomposition

Chapter 3

A Simple Scalable Video Codec Based on MCTF and JPEG2000

This chapter documents the development of a simple scalable video codec based on the integration of motion-compensated temporal filtering (MCTF) and JPEG2000. MCTF is used for temporal wavelet decomposition, and JPEG2000 is used to encode all the T-bands generated by MCTF. By exploring the unique features of MCTF and JPEG2000, the proposed simple codec not only achieves good R-D performance but also has the properties of high scalability and error resilience. The main contributions of the developed codec lie in its simple architecture and the optimum bit allocation across the T-bands based on the quality layer concept of JPEG2000.

3.1 Structure of the Developed Scalable Video Codec

Figure 3.1 shows the structure of the developed encoder. The decoding process simply follows an inverse procedure. The entire encoding process consists of three main components: MCTF, spatial coding and optimal truncation (i.e. optimal rate/bit allocation). In particular, each color component (YUV) of the original frames F_k is first filtered using MCTF with 5/3 wavelet (see Figure 2.10). MCTF is applied iteratively to the set of low-pass T-bands in order to provide multiple frame rates in the final scalable

Figure 3.1: The encoding structure

bitstream. MCTF generates motion vectors and many T-bands. Each T-band can be treated as an individual image. Then, JPEG2000 is used to encode these T-bands with multiple quality layers, each of which has a R-D value. After removing those non-feasible truncation points (during the preprocessing of R-D information), optimal truncation is performed to reach the given target bit rate. The final video bitstream consists of the motion vector (MV) information generated by MCTF and the JPEG2000 bitstream for each T-band. Because all the code blocks of all the T-bands are independently encoded, a feature of JPEG2000, the final bit stream is highly error resilient. In the following, these components are introduced in details.

Figure 3.2 shows typical probability density of two T-bands coefficients (Y component)¹. It could be seen from the figure that the LLLL T-band coefficients mainly spans from 0 to 256 (also true for the U, V components) while the LLH T-band spans mainly

¹Results are generated by applying four levels of temporal decomposition to first 128 frames of the *Foreman* sequence (CIF)

Figure 3.2: Probability density of two T-bands of the Y component

around 0 (also true for all the other high T-bands, i.e. LLLH, LH, and H T-bands). The motion vectors generated at this step will be losslessly coded.

3.2 2D Decomposition and Spatial Coding

The T-bands generated could be treated as individual images. There are various ways to compressed it. JPEG2000 [6] is chosen because it is highly efficient and scalable. In JPEG2000, for 8-byte unsigned (non-negative) binary image, an offset of -128 is added to each pixel value before 2D DWT decomposition so that the samples have a signed representation in the range $[-128, 127]$. The motivation for the offset is that almost all of the T-band samples produced by the DWT involve high-pass filtering and hence have a symmetric distribution about 0. As mentioned above (Figure 3.2), the temporal LLLL T-band appears like a 8-bytes unsigned image while the high T-band is more like a signed image. Therefore, for each high T-band coefficients, the -128 offset will be not applied before the 2D wavelet decomposition.

For non-scalable video coding, quantization step size directly affects the final bit rate and distortion. However, for the quality scalable coding of wavelet coefficients, it is different. Typically some embedded coding method (e.g. bit plane coding as in

JPEG2000) is used to encode the wavelet coefficients. The quantization step has little effect in terms of R-D optimization. However, it must be small enough to allow high quality encoding.

In JPEG2000, quality scalability is achieved through the introduction of quality layers as described in Section 2.4.1. In the developed system, for the base (lowest pass) T-band, around 30 quality layers are used for the each component (range from 0.02 bit per pixel, i.e. 0.02 bpp, to 1.5bpp for Y component, and from 0.005bpp to 0.5bpp for UV component), and for high pass T-bands, around 20 quality layers are used (ranges from 0.005bpp to 1bpp for Y component, and from 0.01 to 0.1 for UV component). The reason to use so many quality layers is to facilitate the optimal bit allocation at different bit rates so that SNR scalability can be easily provided in the final video bitstream. For each quality layer L of each T-band B of each color component C , the corresponding bit rate $r_{L,B,C}$ and its corresponding distortion $D_{L,B,C}$ (compared with uncompressed T-band, measured in mean square error) are recorded.

3.3 Rate control: optimum truncation across T-bands

3.3.1 Theoretical Analysis of MCTF

The result of temporal filtering is the same number of T-bands of different levels. For example, for three levels of MCTF, the four kinds of T-bands are LLL, LLH, LH, and H. Obviously the distortion associated with different T-bands will contribute differently to distortion in the reconstructed frames. Typically, this difference is quantified by analyzing the synthesis filter gain for inverse MCTF. According to the lifting scheme stated in equation (Eq. 2.3 and (Eq. 2.4) (see Figure 2.10), to reconstruct original video frames (i.e. A-frames and B-frames)²:

$$A_k = L_k - \frac{1}{4}H_k - \frac{1}{4}H_{k-1} \quad (\text{Eq. 3.1})$$

²The motion vectors are omitted here

and

$$\begin{aligned} B_k &= H_k + \frac{1}{2}A_k + \frac{1}{2}A_{k+1} \\ &= \frac{3}{4}H_k + \frac{1}{2}L_k + \frac{1}{2}L_{k+1} - \frac{1}{8}H_{k-1} - \frac{1}{8}H_{k+1} \end{aligned} \quad (\text{Eq. 3.2})$$

Conventionally, it is assumed that *the distortion in each encoded T-band is zero-mean additive white noise*. Then if the error variance of pixels in reconstructed T-band F is $\sigma_{\epsilon_F}^2$, then distortion (i.e. sum of squared errors) in reconstructed frames (Eq. 3.1 and Eq. 3.2) is estimated as follows:

$$\sigma_{\epsilon_{A_k}}^2 = \sigma_{\epsilon_{L_k}}^2 + \frac{1}{16}\sigma_{\epsilon_{H_k}}^2 + \frac{1}{16}\sigma_{\epsilon_{H_{k-1}}}^2 \quad (\text{Eq. 3.3})$$

and

$$\sigma_{\epsilon_{B_k}}^2 = \frac{9}{16}\sigma_{\epsilon_{H_k}}^2 + \frac{1}{4}\sigma_{\epsilon_{L_k}}^2 + \frac{1}{4}\sigma_{\epsilon_{L_{k+1}}}^2 + \frac{1}{64}\sigma_{\epsilon_{H_{k-1}}}^2 + \frac{1}{64}\sigma_{\epsilon_{H_{k+1}}}^2 \quad (\text{Eq. 3.4})$$

The equations (Eq. 3.3) and (Eq. 3.4) show the distortion relation between the original frames and the T-bands. This applies for each level of MCTF. Theoretically, MCTF divides the original frames into different frequency subbands with good frequency selection. T-bands of the same level (e.g., HH, LH and L etc.) can be thought to be in the same frequency range. The distortion caused by T-band coding will contribute differently to the total distortion in the final decoded video depends on the level of the T-band and the filtering coefficient used for MCTF. We use the concept of “energy gain” G_{Band} to highlight this difference [21, 29]. The contribution of distortion $\sigma_{\epsilon_{Band}}^2$ in T-band $Band$ to the total distortion in the reconstructed video is $G_{Band}\sigma_{\epsilon_{Band}}^2$. For example, from (Eq. 3.3) and (Eq. 3.4), the distortion $\sigma_{\epsilon_{L_k}}^2$ in the T-band L_k , will cause the distortion in frame A_k , B_{k-1} and B_k to increase by $\sigma_{\epsilon_{L_k}}^2$, $\frac{1}{4}\sigma_{\epsilon_{L_k}}^2$ and $\frac{1}{4}\sigma_{\epsilon_{L_k}}^2$ respectively. Therefore, the contribution of the distortion $\sigma_{\epsilon_{L_k}}^2$ in the T-band L_k to the total distortion in the reconstructed video is $\sigma_{\epsilon_{L_k}}^2 + \frac{1}{4}\sigma_{\epsilon_{L_k}}^2 + \frac{1}{4}\sigma_{\epsilon_{L_k}}^2 = 1.5\sigma_{\epsilon_{L_k}}^2$. In other words, energy gain for L band is

$$G_L = 1 + \frac{1}{4} + \frac{1}{4} = 1.5. \quad (\text{Eq. 3.5})$$

Similarly, we can find the energy gain for H T-band is

$$G_H = \frac{9}{16} + \frac{1}{16} + \frac{1}{16} + \frac{1}{64} + \frac{1}{64} = 23/32 \approx 0.72. \quad (\text{Eq. 3.6})$$

For the case of two levels of MTCF, the same MCTF process (i.e., the same 5/3 filter coefficients) is applied to L T-bands to generate LH and LL T-bands. Therefore, we have $G_{LH} = G_L \times G_H \approx 1.08$ and $G_{LL} = G_L \times G_L \approx 2.25$. For four levels of MCTF with 5/3 filter, the energy gains for LLLL, LLLH, LLH, LH, and H T-bands are then $\{G_L^4, G_L^3 G_H, G_L^2 G_H, G_{LH}, G_H\} = \{5.06, 2.43, 1.62, 1.08, 0.72\}$ respectively.

3.3.2 Bit Allocation Optimization Criteria

In the developed codec, after encoding each T-bands using JPEG2000, next step is to truncate each T-band JPEG2000 stream (for each component) to achieve a target bit-rate R . The problem of optimal bit allocation can be summarized as: given a target bit rate R , how to optimally truncate each T-band JPEG2000 stream so that the overall distortion can be minimized? Here, only quality layers of JPEG2000 streams are considered as the available truncation points although JPEG2000 provides even finer truncation points. This is due to the way how JPEG2000 organizes its bit stream. The basic building block of JPEG2000 is called “packet”, which is formed by the same quality layers from a group of code-blocks, called “precinct”. Each “packet” contains its own header and independent of each other[6]. Truncating only at quality layers of each T-band JPEG2000 stream is straightforward and also helps to maintain the error resilience feature of JPEG2000 streams.

Note that each T-bands JPEG2000 stream contains a number of quality layers (as available truncation points) whose corresponding $r_{L,B,C}$ and $D_{L,B,C}$ are known. The task here to find which quality layer $L_{B,C}$ to choose from each T-band so that the distortion is minimized. In this way, the optimal bit truncation problem could be formulated as

$$D_{total} = \sum_C \sum_B W_C \cdot G_B \cdot D_{L_{B,C},B,C}, \quad (\text{Eq. 3.7})$$

Figure 3.3: Convex hull interpretation

subject to

$$R_{total} = \sum_C \sum_B r_{L_B, C, B, C} < R, \quad (\text{Eq. 3.8})$$

where G_B is the energy gain associated with the T-band B, which is determined by the filter coefficients and the level of MCTF that the T-band belongs to as described above. W_C is the weight ($0 < W_C \leq 1$) for each color component if we want to assign the YUV components with different priority. At present, no special weight is given, i.e. $W_C = 1$.

With the introduction of Lagrange multiplier $\lambda > 0$, the problem is simplified as finding optimum λ to minimize the Lagrange cost J ,

$$J(\lambda) = D_{total}(\lambda) + \lambda \cdot R_{total}(\lambda), \quad (\text{Eq. 3.9})$$

If we can find a value of λ such that the chosen quality layer (truncation points) of each T-band which minimizes (Eq. 3.9) also satisfied $R_{total} = R$, then this set of quality layers to be truncated will be optimal. Because we have a discrete number of quality layers, we only need to find the smallest value of λ subject to (Eq. 3.8).

3.3.3 Preprocessing RD information

Not all quality layers of each T-band are feasible truncation points because the rate distortion curve of the JPEG2000 stream is not a strict convex curve. Figure 3.3 illustrates

this. This encoding block will try to remove those unfeasible truncation points. The second task performed here is to calculate the slope of the RD curve at each truncation point. The slope for truncation point h of T-band B is calculated using the formula

$$Slope_B^h = W_C \cdot G_B \cdot \frac{D_B^{h-1} - D_B^h}{r_B^h - r_B^{h-1}}. \quad (\text{Eq. 3.10})$$

The slope and rate information are required later on for optimum truncation across T-bands.

3.3.4 Finding optimum truncation points

The task here to find optimum λ for a given bit-rate R using the rate and slope information from previous encoding block. We apply a binary searching method as illustrated in Figure 3.4. Starting with an initial value, the search algorithm could quickly find the optimum λ , which will then determines the selection of truncation points from each T-band JPEG2000 stream to achieve the target bit rate in the final bit stream R . In order to provide SNR scalability in the final bitstream, this optimum rate control method will be applied to a sequence of target bit-rates $R_1, R_2, \dots, R_n$.

3.4 Experimental Results

The proposed codec have been tried on three standard test sequences, Akiyo, Foreman and Mobile&calendar. Figure 3.6 shows some sample frames decoded. Figure 3.5 compares the rate distortion performance with H.264/AVC. ³ It could be seen that for low motion sequence (Akiyo) and high texture sequence (Mobile&calendar), the performance of the proposed codec is very close to H.264. However, for the Foreman sequence which has higher motion, the proposed codec performs worse than H.264. It could be lowered

³Coding parameters for H.264: JM 9.8, Main profile, structure: IBPBP..., I frame refreshing rate: 16 frames. Coding details for the proposed codec: first 128 frames, 4 levels of MCTF, i.e. five kinds of T-bands, LLLL, LLLH, LLH, LH, and H.

CHAPTER 3. A SIMPLE SCALABLE VIDEO CODEC BASED ON MCTF AND JPEG2000

Figure 3.4: Finding the optimum truncation points

by 2dB compared to H.264 with five reference frames. There could be two reasons: 1). H.264 allows multiple reference frames at each motion estimation direction, where in our experiment, MCTF use only one reference frame. 2). Block-based motion estimation used within MCTF causes many block anti-facts within some T-bands, especially temporal high bands. These block anti-facts causes many vertical and horizontal edges inside high bands which will be encoded using JPEG2000, and hence introduce performance drop. Because this edges are un-natural and JPEG2000's RD performance is not so good when encoding unnatural images. In the future, these effects caused by these two problems could reduced by the following: 1). Introducing multiple reference with MCTF, as in [30]. 2). The H.264/AVC [12] de-blocking filter could be applied the T-bands before encoding using JPEG2000.

Although the proposed codec performs worse than H.264 for some sequences, it still has advantages over H.264. With the proposed codec, to achieve different bit rates, we

CHAPTER 3. A SIMPLE SCALABLE VIDEO CODEC BASED ON MCTF AND JPEG2000

Figure 3.5: RD performance in terms of PSNR

CHAPTER 3. A SIMPLE SCALABLE VIDEO CODEC BASED ON MCTF AND JPEG2000

Figure 3.6: Sample reconstructed frames

only need to truncate the encoded bitstream accordingly rather than repeatedly encoding the entire sequence for each required bit rate as with H.264. Furthermore, with the proposed codec, a video stream with lower spatial resolution (supported by 2D wavelet in JPEG2000) or lower temporal frame rate (supported by MCTF) can be easily generated by truncating the bitstream appropriately. For the case of H.264, the only way to achieve this scalability is to decode the video sequence and re-encode with different spatial resolution or temporal rate. If the simple codec proposed here gives such a good performance in coding efficiency and scalability, we can predict that more sophisticated methods based on 3D wavelet will definitely give better results.

Besides its coding efficiency and scalability, the developed codec also provides many other useful features since it is based on JPEG2000. One of them is region of interest (ROI) encoding [6, 24], which allows certain region of an image to be encoded with greater fidelity (quality) than the rest of the image. The basic premise behind ROI is to encode wavelet coefficients that contribute to the ROI with higher precision than the others by shifting the position of bit planes. This feature has been tried on the sequence Akiyo. The head region is encoded with higher priority. Figure 3.7 illustrates its effect. It can be seen that with ROI coding the head appears clearer while the rest becomes more blurred.

3.5 Summary

In short, a simple scalable video codec based on 3D wavelet was proposed. MCTF is applied for temporal filtering, while the 2D spatial wavelet decomposition and coding is based on JPEG2000. With the “quality layer” concept in JPEP2000, a simple optimum rate control mechanism is used to generate a video bitstream supporting multiple bit rates (i.e., SNR scalability). The experimental results show that the codec’s performance is comparable with state-of-the-art coding standard H.264/AVC, especially for low motion

CHAPTER 3. A SIMPLE SCALABLE VIDEO CODEC BASED ON MCTF AND JPEG2000

Figure 3.7: ROI coding (left: 23kbps without ROI. right: 23kbps with ROI coding)

or high texture video sequences. The experimental results show the great potential of the 3D wavelet based coding methods.

Although several codecs based on similar integrations have been proposed in the literature [28, 31, 30, 32], most of them have complex coding structure in order to pursue superior R-D performance against the state-of-the-art H.264/AVC. The main contributions of the proposed codec lie in its simple architecture and the optimum truncation across T-bands to generate different bit rate. More over, many features in JPEG2000, such as ROI coding and error resilience, can be easily utilized with the proposed codec. Furthermore, the proposed simple codec could serve as a good framework for the study of other important issues related to 3D wavelet based coding, such as constant quality aimed bit allocation and the transportation of video stream over networks, which will be addressed in the rest of this thesis.

Chapter 4

Constant Quality Aimed Bit Allocation for 3D Wavelet Based Scalable Video Coding

As stated earlier, 3D wavelet that extends 2D wavelet transform to motion pictures has received much attention recently. The current trend of 3D wavelet codecs [2, 3, 4] is to apply MCTF for temporal decomposition and 2D wavelet for further spatial decomposition. For MCTF based codecs, one fundamental question is that, given a bit budget, how to allocate bits to each T-band so that certain degree of constant quality can be achieved. Although the problem of constant quality aimed bit allocation has been well studied for conventional hybrid codecs [33, 34], it is still an open question for MCTF based codecs. For conventional hybrid video coding method such as MPEG-4 or H.264, which encodes video sequence frame by frame, it is easy to control the quality of individual frames by adjusting quantization parameters or selecting enhancement layer bit planes as in MPEG-4 FGS (see Section 2.3.2). However, for MCTF based 3D wavelet coded, it is very different. Recently, we have seen some research work [35, 31, 36, 37] looking into this issue. In the popular MC-EZBC codec [35], to ensure constant quality across GOPs, it stops bit plane scanning of all the Groups of pictures (GOPs) at the same fractional bit plane. However, this could only help achieve a similar distortion among T-bands, which

CHAPTER 4. CONSTANT QUALITY AIMED BIT ALLOCATION FOR 3D WAVELET BASED SVC

does not lead to constant quality in reconstructed frames. This is because the distortion in one T-band could propagate unevenly into the reconstructed frames in a way that is very hard to model mathematically with conventional analysis. This uneven distribution of distortion among reconstructed frames makes it difficult to decide the bit allocation for each T-band in order to achieve a constant quality for reconstructed video while maintaining the total bits within a given constraint. In [31, 36], in order to smooth PSNR behavior, an optimized quantization step is obtained by analyzing the motion behavior during the temporal filtering. In [37], the authors propose to use an adaptive update step for the temporal filtering. Compared with conventional implementation, these methods indeed help reducing the serious PSNR fluctuation in reconstructed frames. However, they did not explicitly address the problem of bit allocation among T-bands.

In this chapter, a novel algorithm on constant quality aimed bit allocation among T-bands is proposed for the applications of adaptive stored video streaming with the scalable video codec presented in Chapter 3. In particular, an approximate mathematical model is proposed to describe the relationship between the T-band distortions and the distortions of the reconstructed frames. Since only stored video is considered in this research, the model parameters can be off-line generated. During the online transmission stage, given the current network bandwidth, the conventional JPEG2000-like optimum truncation (refer to Section 3.3) is performed first. After that, a two-step operation is used for reducing the PSNR fluctuation, where the basic idea is to adjust the energy gains to compensate the different contributions from different types of T-bands and then more or less equally distribute distortion among the T-bands at the same level. Experimental results show that the proposed scheme can greatly reduce PSNR fluctuation with only slight degradation in average PSNR. First, the problem with the conventional theoretical analysis for MCTF will be reviewed before the description of the proposed algorithm.

Figure 4.1: Distortion relationship between frame L_k and frame B_k .

4.1 Problem Statement

Section 3.3.1 presents the conventional theoretical analysis for MCTF. The equations (Eq. 3.3) and (Eq. 3.4) show the distortion relationship between reconstructed frames and each T-band. Based on them, the energy gain associate with each level of T-bands is estimated and then used to guide the bit allocation for the optimum Rate-Distortion performance (i.e. JPEG2000-like optimum truncation). However, the assumption that distortions in T-bands are white noise is obviously not accurate. In addition, due to the imperfect motion compensation process, the correlation among T-bands can not be ignore [29], which makes it inaccurate to estimate the reconstructed frame distortion using (Eq. 3.3) and (Eq. 3.4).

To verify the accuracy of the conventional theoretical analysis, some experiments are done. For one level MCTF (see Figure 2.10), according to equation (Eq. 3.4), a change c in the distortion of T-band L_k (either increase or decrease) will cause the corresponding distortion change of $0.25c$ in the reconstructed frame B_k . Figure 4.1 shows the experimental results using the developed JPEG2000 based codec (refer to Chapter 3) with two CIF sequences *Table-tennis* and *Stefan*. The X-axis shows distortion change in T-

CHAPTER 4. CONSTANT QUALITY AIMED BIT ALLOCATION FOR 3D WAVELET BASED SVC

Figure 4.2: PSNR estimated with conventional MCTF distortion analysis.

band L_k while the Y-axis shows the corresponding distortion change in the reconstructed frame B_k . Obviously, it could be seen that the coefficient $1/4$ in equation (Eq. 3.4) is not accurate at all. Similarly, the coefficients $1/16$, $9/16$ and $1/4$ in (Eq. 3.3) and (Eq. 3.4) are not correct either. Fig. 4.2 shows the PSNR at 800Kbps with three levels of MCTF for Stefan_CIF sequence. The estimated PSNR based on Eqs. (Eq. 3.3) and (Eq. 3.4) is also plotted in Fig. 4.2. The difference in the actual PSNR and the estimated PSNR is obvious. The estimated average PSNR is 33.57dB while the actual average PSNR is 32.61dB. With experiments it is also found that the additive property of MSE in (Eq. 3.3) and (Eq. 3.4) also does not apply to MCTF because of the assumption of un-correlation (see Section 3.3.1) is incorrect. Due to these inaccuracies in (Eq. 3.3) and (Eq. 3.4), the resulted energy gains of the T-bands have many side-effects. One of them is the PSNR fluctuation in reconstructed frames. Figure 4.3 shows an example with a popular 3D wavelet based codec MC-EZBC using *Stefan* CIF sequence. Two levels of MCTF are applied with the first 256 frames. Though the filter used by MC-EZBC [35] is Harr filter¹. The underline principle is still the same. From the figure, it could be noticed that the frames A0, A2, ..., have the highest PSNR. This is because they are directly related

¹Actually, in MC-EZBC, the filter used is $(\sqrt{2}/2, \sqrt{2}/2)$ for L band, and $(\sqrt{2}/2, -\sqrt{2}/2)$ for H band, which results in the same energy gain for L band and H band theoretically. However, compared with traditional Harr filter that uses $(1/2, 1/2)$ for L band, $(1, -1)$ for high band, the L band is effected increased by $(\sqrt{2})$ time while the H band is reduce by $(\sqrt{2})$ times

CHAPTER 4. CONSTANT QUALITY AIMED BIT ALLOCATION FOR 3D WAVELET BASED SVC

Figure 4.3: PSNR with MC-EZBC for Stefan_CIF at 1.2Mbps with two levels of MCTF.

to the lowest T-band LL bands, which are given higher priority due to higher effective energy gain associated with these T-bands. Similarly, frames A1, A3, ..., have higher PSNR than frames B0, B1, With the developed JPEG2000 based codec, a similar phenomenon is observed as illustrated in Figure 4.4.

Figure 4.4: PSNR with the developed JPEG2000 based codec for Stefan_CIF at 1Mbps with two levels of MCTF.

Figure 4.5 shows the actual MSE value for each T-band optimum bit allocation for the JPEG2000 based codec. Due to the optimum truncation method, the fluctuation of MSE among T-Bands is quite severe. There could be two reasons for the PSNR fluctuation in the reconstructed frames. One obviously reason is the fluctuation of MSE

Figure 4.5: Typical MSE for each T-band with the JPEG2000-like optimum bit allocation

for the T-bands at the same level (e.g. all LL bands, LH bands). This is mainly due to different RD curve for each T-band and the use of Lagrange-multiplier based optimum allocation method that only considers the slope along the RD curve rather than the actual distortion value (i.e. aims to minimize average distortion instead of quality difference among reconstructed frames). The other reason is the MSE change across the T-bands at different levels, resulted by the inaccurate energy gains used to guide the allocation. This is also the reason why frames A0, A2, A4, ..., have higher PSNR than the other frames, as illustrated in Figure 4.3 and Figure 4.4.

4.2 Constant Quality Aimed Bit Allocation

Ideally, a mathematical model is needed in order to achieve constant quality bit allocation. However, it is non-trivial to develop such a model. First, due to the complex motion compensation process, distortion at different regions of a T-band propagates differently to reconstructed frames. A precise model would require distortion analysis at pixel or block level. Then it will be necessary to consider finer bit allocation scheme, such as bit allocation among small code-blocks of 3D wavelet coefficients. In addition, the unknown correlation among T-bands further complicates the distortion relationship

between reconstructed frames and T-bands, and hence makes it difficult to estimate the distortion of individual frames. Here, a simple approach is proposed to reduce quality fluctuation.

As showed above, one reason for PSNR fluctuation is the uneven MSE distribution among the T-bands at the same level. Intuitively, the distortion of the T-bands at the same level should be close to each other since they contribute more or less equally to the distortion of the reconstructed frames. Therefore, a straight-forward method is to re-adjust the bit allocation optimum truncation so that their MSE value are close. In particular, in the JPEG2000 based codec, for a T-band level l , the average MSE $D_{avg}(l)$ of all the T-bands at this level is calculated. After that, for each T-band at level l , the selected truncation point is adjusted to a neighboring feasible truncation point that has a closer MSE to $D_{avg}(l)$, i.e. to select a neighboring truncation point which is close to the one selected by Lagrange optimum bit allocation and also has a MSE value that is close to the average MSE value of all the T-bands at the same level. Figure 4.6(a) shows the effect of this simple adjustment on the Stefan_CIF sequence at 1Mbps with two levels of MCTF. Figure 4.6(b) shows the resulted PSNR compared with the original PSNR. From the experimental results, it could be seen that this smoothing of MSE distribution indeed helps to reduce the PSNR fluctuation within a same class of frames (such as A0, A2, A4, ...). Since this adjustment is done towards the average MSE of the T-bands at the same level, the total bit rate does not change much (usually less than 5%). Of course, the average distortion will be dropped since the new bit allocation is not targeted to minimize the average distortion any more. However, from experiments, it is observed that the drop in average PSNR is usually less than 0.3dB. The reasons are (1): the original “optimum bit distortion” as stated in Chapter3 is not really optimum because of the use inaccurate energy gain, and (2) this adjustment only smooths the MSE behaviors of the T-bands at the same level and it does not change the average MSE for them.

CHAPTER 4. CONSTANT QUALITY AIMED BIT ALLOCATION FOR 3D WAVELET BASED SVC

(a) Smoothing MSE of T-bands

(b) PSNR result after the smoothing of MSE of T-bands

Figure 4.6: Smoothing MSE of the T-bands at the same level for Stefan_CIF at 1Mbps

Although the previous adjustment does help reduce the fluctuation of PSNR within the same class of frames, there still exists large PSNR difference among different classes of frames. For example, the group of frames A0, A2, ..., still have the highest PSNR in Figure 4.6(b). This is because this group of frames are directly related to the lowest T-bands, i.e. LL bands, which has the highest energy gain. To change this, it is needed to change the energy gain for the optimum bit allocation. The problem left is how to change. The theoretical analysis in (Eq. 3.3) and (Eq. 3.4) has been proven not accurate

CHAPTER 4. CONSTANT QUALITY AIMED BIT ALLOCATION FOR 3D WAVELET BASED SVC

by experiments. One way to solve this problem is to empirically modify the energy gains until the average PSNR difference among different classes of the reconstructed frames becomes very small. However, to search the energy gains for each different bit rate is intolerable for adaptive online transmission. Therefore, another way to find the appropriate energy gains is proposed here. Specifically, the theoretical model in (Eq. 3.3) and (Eq. 3.4) are modified into an empirical method. Let us consider one level of MCTF (see Figure 2.10). From (Eq. 3.1) and (Eq. 3.2), the following distortion relationship between reconstructed frames and the T-bands could be assumed:

$$\sigma_{\epsilon_{A_k}}^2 = \alpha_{A_k} \cdot \sigma_{\epsilon_{L_k}}^2 + \beta_{A_k} \cdot \frac{1}{16} \sigma_{\epsilon_{H_k}}^2 + \gamma_{A_k} \cdot \frac{1}{16} \sigma_{\epsilon_{H_{k-1}}}^2, \quad (\text{Eq. 4.1})$$

and

$$\sigma_{\epsilon_{B_k}}^2 = \alpha_{B_k} \cdot \sigma_{\epsilon_{L_k}}^2 + \beta_{B_k} \cdot \frac{1}{16} \sigma_{\epsilon_{H_k}}^2 + \gamma_{B_k} \cdot \frac{1}{16} \sigma_{\epsilon_{H_{k-1}}}^2, \quad (\text{Eq. 4.2})$$

where $\alpha_{A_k}, \beta_{A_k}, \gamma_{A_k}, \alpha_{B_k}, \beta_{B_k}, \gamma_{B_k}$ are unknown positive scaling factors. If these values are known, the distortion in the reconstructed frames using the distortion information of each T-band could then be estimated. These unknown parameters are computed off-line with multiple passes of encoding and decoding and then recorded for future usage². Clearly, this modified model is also inaccurate to estimate the distortions for some individual frames since we still ignore the correlation among T-bands. Nevertheless, for most of the frames, the estimation is accurate. It is also sufficient to estimate the average distortion among one class of frames (e.g. those directly related to low-pass T-band). Fig. 4.7 shows the new estimated PSNR for Stefan_CIF at 800kbps with three levels of MCTF and its difference with the actual PSNR³. Compared with estimation result of theoretical model (i.e. Eqs. (Eq. 3.3) and (Eq. 3.4)) in Fig. 4.2, the new model gives much more accurate

²Although recoding of this coefficients would take some storage space, however it is negligible compared large video stream. Besides, these coefficients only needs needs to stored at source and will not be transmitted to any video clients

³At present, all the model parameters are estimated based on the result from three sets of encoding and decoding passes.

CHAPTER 4. CONSTANT QUALITY AIMED BIT ALLOCATION FOR 3D WAVELET BASED SVC

estimation of reconstructed frames's PSNR based on the distortion of T-bands. With the same sequence, Fig. 4.8 compares the estimated average PSNR with the theoretical model at different bit rate with different levels of MCTF applied. Again, the results shows the accuracy improvement.

(a) Estimated PSNR with the proposed model

(b) Estimation error with the proposed model

Figure 4.7: Estimating PSNR with the proposed model (Stefan_CIF at 800kbps)

(a)Stefan_CIF with three levels of MCTF

(b)Stefan_CIF with two levels of MCTF

Figure 4.8: Estimating average PSNR with different models

With the proposed model which presents a much more accurate distortion relationship between T-bands and decoded frames than conventional theoretical model, there are many ways to utilize it for MCTF based SVC. For example, as mentioned earlier, for constant quality aimed video adaptation, we could easily try different energy gains values to find an appropriate set of energy gains so that PSNR fluctuation among different groups of frames could be reduced. Based on the proposed model, we proposed a novel approach for constant quality aimed bit allocation among T-bands, which can be summarized as follows.

- Step 1. During the JPEG2000-like optimal truncation, modify the energy gains so that the average estimated distortions for different classes of frames using Eqs. (Eq. 4.1) and (Eq. 4.2) are close to each other. With this, we bypass complex distortion analysis between T-bands and reconstructed frames.
- Step 2. Readjust the bit allocation to smooth the MSE distribution among the T-bands at the same level. This helps to take the distortion value into account rather than only considering the slope of the rate-distortion curve in typical Lagrange-based optimum bit allocation.

4.3 Experimental Results

The proposed algorithm has been tried on the sequences *Stefan_CIF*, *Foreman_CIF* and *Table_tennis_CIF* with two levels of MCTF (only the first 256 frames are used here). Figure 4.9 shows the result for Stefan. Comparing Figure 4.9(a) with Figure. 4.6(a), it could be noticed that the average MSE for the LL T-bands is increased while the average MSE for the H T-bands is reduced. The effect of this is the reduced PSNR difference between the three classes of frames. This is due to the change of energy gains for different T-bands. The proposed method could achieve a very smooth PSNR in the

CHAPTER 4. CONSTANT QUALITY AIMED BIT ALLOCATION FOR 3D WAVELET BASED SVC

(a) Smooth the MSE distribution for T-bands after modifying the energy gain

(b) PSNR after applying the constant quality adaptation method

Figure 4.9: Constant quality aimed bit allocation for Stefan_CIF at 1Mbps

Figure 4.10: Experimental results for Foreman_CIF at 400kbps

reconstructed frames compared with the original Lagrange JPEG2000-like optimum bit allocation. Specifically, the PSNR variance is reduced from 0.86 to 0.12 while the average PSNR is dropped about 0.4 dB, which is acceptable in practice. The drop in average PSNR is the kind of trade-off that has to be paid for constant quality adaptation, which is also the case for other video codec like MPEG-4 etc. The results with sequence Foreman and Table_tennis (Figure 4.10 and Figure 4.11) show similar performance. Table 4.1 summarizes the results for the three sequences at different bit rates. It could be seen that the variance, which is used here to indicate the fluctuation in the PSNR value, has been dropped significantly with the proposed methods while the average PSNR for the reconstructed frames does not drop much, which implies the effectiveness of the proposed algorithm. In general, the proposed scheme can reduce PSNR variance 3 ~ 10 times while average PSNR degradation is not more than 0.6 dB (see Table 4.1). This degradation in average PSNR will not cause much problem because the individual frames whose PSNR get dropped are those with high PSNR originally. Their quality gets dropped so that more bits can be allocated to those who have low PSNR originally.

CHAPTER 4. CONSTANT QUALITY AIMED BIT ALLOCATION FOR 3D WAVELET BASED SVC

Figure 4.11: Experimental results for Table_tennis_CIF at 600kbps

Video (CIF) Sequence	Bit Allocation Scheme	Average PSNR (dB)	PSNR Variance
Stefan (800kbps)	JPEG2000-like	31.5	0.78
	Proposed	31.2	0.13
Stefan (1.0Mbps)	JPEG2000-like	32.7	0.86
	Proposed	32.3	0.12
Stefan (1.20Mbps)	JPEG2000-like	34.0	1.62
	Proposed	33.4	0.23
Foreman (250kbps)	JPEG2000-like	32.2	1.33
	Proposed	31.9	0.08
Foreman (400kbps)	JPEG2000-like	34.1	0.84
	Proposed	33.7	0.08
Foreman (500kbps)	JPEG2000-like	35.0	0.96
	Proposed	34.7	0.11
Table tennis (300kbps)	JPEG2000-like	30.6	0.68
	Proposed	30.2	0.16
Table tennis (450kbps)	JPEG2000-like	32.4	0.55
	Proposed	32.0	0.17
Table tennis (600kbps)	JPEG2000-like	33.7	0.62
	Proposed	33.1	0.20

Table 4.1: Comparison of variance and average of the PSNR between the proposed constant quality aimed bit allocation scheme and the JPEG2000-like bit allocation scheme.

4.4 Summary

To summarize, this chapter proposed a novel approach for constant quality aimed bit allocation among T-bands for 3D wavelet based scalable video codec. The problem with the conventional MCTF analysis and the JPEG2000-like bit allocation is analyzed. After observing the problem of the significant PSNR fluctuation, the causes in the conventional bit allocation are pointed out. Based on the analysis, here it has been proposed to adjust the energy gains of T-bands at different levels in order to make the distortions of different classes of reconstructed frames close to each other. Furthermore, by smoothing the MSE allocation for T-bands at the same level, it have further reduced the PSNR fluctuation for reconstructed frames in the same class. Experimental results have demonstrated the effectiveness of the proposed method.

Chapter 5

Application of 3D Wavelet Based Scalable Video Coding for Video Streaming

Scalable video coding schemes are intended to encode the video signal once, but enable decoding from partial streams depending on the specific bit rate and resolution required by a certain application. This enabled a simple and flexible solution for transmission over heterogeneous networks, additionally providing adaptability for bandwidth variations and error conditions. Based on the 3D wavelet SVC developed in Chapter 3, this chapter shows some examples of utilizing scalable video coding on the application of stored video streaming.

5.1 An End-to-end Video Streaming System for Best-effort Internet

This section documents the development of an end-to-end streaming system for transporting stored video on the present best-effort IP network. Note that here only uni-cast point-to-point transport of stored video (pre-encoded) is considered.

5.1.1 Background

The present packet-switched Internet was designed to interconnect a variety of different networks by the using of routers. The original design philosophy concentrates on the robustness of this interconnection in case of network failure and the optimization of the average throughput with little consideration on the timeliness of data delivery or bandwidth reservation or any other quality of service (QoS) guarantees. In other words, the current version of Internet protocol, IPv4, provides only so called best-effort service. Network bandwidth, packet delay could change at any time. In addition, no any loss guarantee is provided in IPv4. IP packets could be discarded by the routers because of congestion which could happen at any time during the transmission (i.e. unpredictable). However, the real-time delivery of video is subject to strict bandwidth, loss and time constraints. Therefore it is a challenging problem to design an efficient video transport system that can maximize the video perceptual quality under varying network conditions, i.e. to make best use of the best-effort Internet.

To address above issues without any modification to the currently deployed Internet, end-system based approaches are commonly utilized [38]. In particular the end systems (i.e. the streaming server and clients) employ control techniques to maximize video quality without any QoS support from the Internet. Such an approach is of particular significance and applicable to both current and future Internet since it does not require the participation of networks. It requires end-to-end applications to determine network status at any given time and act accordingly to maximize transported video quality. Therefore it is also called application-layer QoS control. The objective is to avoid congestion and maximize video quality in the presence of packet loss.

Ideally the streaming server needs to send the video stream at the available bandwidth so that the best-possible quality could be achieved at the client side. However, the

CHAPTER 5. APPLICATION OF 3D WAVELET BASED SVC FOR VIDEO STREAMING

available bandwidth not only is not known a priori but also varies with time. Then how will the server know the network conditions and send the video with appropriate bit rate? Since the network (i.e. IP routers) does not convey any useful information regarding the network conditions, only the end systems could do that. Obviously the streaming client in the receiving terminal is at the best position to estimate the status of the network by monitoring the packet loss rate or variation in packet delay [39, 40]. Many end-to-end control mechanisms have been proposed for the server to adapt to network status change based on the feedback information from video receiver [38] based. There are usually two ways, namely, the probe-based approach and model-based approach.

Probe-based approach: this approach implicitly estimates the available bandwidth through probing experiments. The server probes for the bandwidth by adjusting video rate in a way that could maintain some QoS requirements, e.g. packet loss ratio p is below a certain threshold p_{th} [38]. There are two ways to adjust the rate: (1) additive increase and multiplicative decrease (AIMD), and (2) multiplicative increase and multiplicative decrease (MIMD). In (1), the rate is modified as follows:

$$r(t+1) = r(t) + a, \text{ if } p < p_{th}, \text{ and } r(t+1) = r(t) \times b, 0 < b < 1 \text{ if otherwise, (Eq. 5.1)}$$

where additive increase rate (AIR) a and multiplicative decrease factor (MDF) b are positive constants. For MIMD, the following equation is used:

$$r(t+1) = r(t) \times b, \text{ where } b > 1 \text{ if } p < p_{th}, \text{ and } 0 < b < 1 \text{ if otherwise. (Eq. 5.2)}$$

It has been shown that a simple AIMD algorithm satisfies sufficient conditions for the convergence to an efficient and fair state regardless of the starting state of a network [41].

Model-based approach: it attempts to estimate the available bandwidth explicitly by using a throughput model of a TCP connection, which can be characterized by [42]:

$$r = \frac{1.22 \times MTU}{RTT \times \sqrt{p}}, \quad (\text{Eq. 5.3})$$

Figure 5.1: The System architecture for the end-to-end streaming system

where r is the throughput of a TCP connection, MTU (maximum transit unit) is the packet size used by the connection and RTT is the round-trip time and p is packet loss ratio experienced by the connection. Using this model, the streaming session could avoid congestion in a similar way to that of TCP connection and compete with other TCP flows in a "friendly" manner [42].

After the streaming server determined the desired video sending rate no matter which approach is used, it has to apply some mechanism to shape the bit stream from the video encoder to enforce that rate, i.e video rate adaptation. For example, for live video, this is often achieved by introducing a rate-adaptive video encoding algorithm to the video encoder as in [39, 43, 44]. For stored video encoded with the developed 3D wavelet based scalable codec, this rate adaptation is relatively straight-forward.

5.1.2 End-to-end System Design

With the above discussion, an end-to-end streaming system is proposed, as showed in Figure 5.1. There are two main data paths on the IP network, namely, the data plane (transporting actual video streams) and the control plane (transporting various control messages, including feedback from the streaming clients).

The main components in the system are Pre-encoded Video, Session Control, Bit Allocation, Network Access Layer and Video Decoder. First, the raw video is pre-encoded with the 3D wavelet based SVC described in Chapter 3. The main function for Session Control at the server side is congestion control based on feedback information from the receiver. A feedback control scheme is applied here to estimate the available network bandwidth. A probe-based approach with a modified AIMD algorithm is chosen for bandwidth estimation (see **Algorithm 1**). It is slightly different from traditional AIMD schemes in term of the value of packet loss threshold and MDF factors. With **Algorithm 1**, the MDF factor is modified based on different value of loss rate detected by the video receiver. This helps the system to quickly adapt to abrupt bandwidth change. The feedback (i.e. packet loss rate) is generated by the Session Control module at the receiver's side, a kind of QoS monitor, which monitors the behavior of the arriving packets, such as packet loss rate and delay. The packets received correctly and timely will be passed to decoder to reconstruct video frames for display. Then, "Bit allocation" is responsible to adapt the video bit rate to the estimated bandwidth (see Section 3.3). The feedback control scheme could be considered here as rate control from the transport perspective while Bit Allocation is from the video compression perspective. Note this AIMD control scheme has been proven to be TCP friendly with simulation in [39] and with actually experiment in [44]. In other words, even though UDP is used for stream transportation, it will not influence the TCP-friendliness in the best-effort Internet.

CHAPTER 5. APPLICATION OF 3D WAVELET BASED SVC FOR VIDEO STREAMING

It is necessary here to point out that the bit allocation in Section 3.3 is actually a variable bit rate (VBR) bitstream with an average bit rate of R since the optimal bit allocation algorithm only minimizes the average distortion under the constraint of a total bit budget. A burst of packets may be generated at one time and passed to the network layer for transmission. Since network layer, especially the MAC layer, is often implemented in firmware with very limited resource including buffer size, a burst of packets may cause unnecessary buffer overflow. As such, the optimal bit allocation is applied to each small time interval $T_{interval}$ independently. In other words, the T-bands in each $T_{interval}$ is allocated a total bandwidth of $R \cdot T_{interval}$. A leaky bucket is further applied to smooth out the bit rates allocated to each T-band. In this way, the application layer at the server will generate a corresponding constant bit rate (CBR) like bit stream to the network layer.

Algorithm 1 Feedback Control Scheme at Application Layer

Initial Rate: IR ; Maximum Rate: $MaxR$; Minimum Rate: $MinR$;
Additive Increase Rate: AIR ; Multiplicative Decrease Factor: MDF ;
Estimated Bandwidth: $R = IR$;
for each feedback received with Packet Loss Ratio P_{loss} **do**
 if $P_{loss} < 0.5\%$ **then**
 $R = \min(MaxR, R + AIR)$;
 else
 if $P_{loss} > 20\%$ **then**
 $MDF = 0.5$;
 else if $P_{loss} > 3\%$ **then**
 $MDF = 0.7$;
 else
 $MDF = 0.8$
 end if
 $R = \max(MinR, R \times MDF)$;
 end if
end for

For the packetized video transmission, T-band streams and MV data need to be assembled into individual network packets (i.e., packetization). In this thesis, it is simply

done by putting different T-band of each color component into different packets and placing motion vectors into the first packet of their associated temporal high bands of the Y color component. The size of a packet is limited by a predefined maximum packet size $Psize$.

The network protocol stack chosen for the architecture is RTP/UDP/IP. TCP is not chosen because the re-transmission mechanism used by TCP introduces unnecessary delay which is unacceptable for real-time video applications. Further more, the window-based congestion control mechanism applied in TCP might cause rapid change in the data sending rate. Therefore UDP is employed as the transport protocol. Since UDP provides an unrealizable and connectionless service, UDP packets might be lost during transmission or received out of sequence. This unreliability could be dealt with by the receiver with the support from the upper layer, i.e. RTP/RTCP. Real-time Transport Protocol (RTP) is an Internet standard protocol designed to provide end-to-end network transport functions suitable for applications transmitting real-time data [45]. Real-time Control Protocol (RTCP) is a companion control protocol with RTP. It is designed to work in conjunction with RTP to provide QoS feedback to the participants of a RTP session. In the end-to-end system proposed, it is used to carry the feedback information from the streaming clients. In other words, RTP is a data transfer protocol while RTCP is a control protocol.

5.1.3 Experimental Results

In order to test whether the above designed streaming system can really adapt the video sending rate to available network bandwidth, a simple peer-to-peer benchmark network configuration shown in Figure 5.2 is employed for the Internet environment. Such a simple network configuration captures the fundamental property of a transport path within the

Figure 5.2: A peer-to-peer network

Internet since there is usually only one bottleneck link (i.e. the one with minimum bandwidth among all the traversing links) between the sender and the receiver.

The network simulation tool used is Network Simulator (NS-2), a discrete event network simulator originally developed at UC Berkeley that simulates variety of computer networks [46]. The standard raw video sequence "table-tennis" is used as the source video. The video¹ is encoded with the developed SVC. Two levels of MCTF are applied for the encoding. The sequence is repeated cyclically during the 270 seconds test run.

The link rate of the bottleneck link varies from 1000 kilo bits per second (kbps) during [0 70) second to 500 kbps during [70, 180) second to 800 kbps after 180 second (see Figure 5.3(b)). The simplest queuing algorithm, "drop tail", is applied at both routers during the simulation. The server follows the feedback control scheme presented in **Algorithm 1** with the following parameter values: $IR = 700\text{kbps}$; $MinR = 200\text{kbps}$; $MaxR = 1000\text{kbps}$; $AIR = 2\text{kbps}$. The packet loss ratio is updated by the receiver every $T_{feedback} = 2$ seconds.

Figure 5.3 shows the experimental result. The network link bandwidth and video bit rate behavior during the 270-sec simulation run are plotted in Figure 5.3(b). It could be seen that the streaming server is able to adjust its output video rate to keep track of the varying available network bandwidth with the feedback control performed at the application layer. Figure 5.3(a) shows the packet loss ratio detected by the client during

¹Frame size: CIF; only the first 256 frames are used for simplicity; $T_{interval}$ mentioned in Section 5.1.2 is set to 32/30 seconds

CHAPTER 5. APPLICATION OF 3D WAVELET BASED SVC FOR VIDEO STREAMING

Figure 5.3: Performance of the end-to-end streaming system

the same simulation run, which is consistent with that showed in Figure 5.3(b). The oscillation in the sending rate is due to binary nature of the feedback control scheme (i.e. decrease rate if loss ratio is larger than a threshold P_{th} , and increase if otherwise). From Figure 5.3(c), it could be seen that video quality improves as bit rate is increased which illustrates the effectiveness of the streaming system. Due to adaptation of video bit rate, the packet loss ratio is kept low most of the time, which makes the PSNR drops gracefully when bandwidth drops. For example, at 70 second when bandwidth suddenly drops to 500kbps, PSNR drops from about 36dB to about 32dB with a short duration of unacceptable PSNR (small than 28dB). One thing to emphasis here is that due to scalable video coding, the adaptation of rate could be done very quickly (i.e. without

much computation) because no re-encoding is needed.

However, due to the nature of system, it is still unavoidable that packets could get lost and the effects of packet loss is not well-controlled. At time 70 second when the PSNR suddenly drops significantly, it still takes time (about 5 seconds) for the server to reduce the rate so that the PSNR goes back to an acceptable level. This is because the server needs to wait for feedback information which is generated by the client every $T_{feedback} = 2$ seconds. Although it could be improved by decreasing $T_{feedback}$, small $T_{feedback}$ might result in inaccurate measurement of packet loss ratio and increase of bandwidth waste for control messages. Even when bandwidth remains at 500kbps during [70, 180) second, occasional packet loss could also cause PSNR to reduce greatly. This is because all packets are treated equally. When some important packets, such as the first packet of a low pass LL T-band which could cause all the other packets of the same band undecodable and hence large PSNR drop. These could be "solved" through some unequal error protection schemes at application layer [47, 48, 49] or packet retransmission [50]. But these methods will greatly increase system complexity at the end systems.

5.2 Video streaming Over QoS Enabled Network, IEEE 802.11e WLAN

From previous section, it could be seen that received video quality can be further improved if the network layer could inform streaming server about any bandwidth change or important packets are better protected (e.g. un-important packets are dropped first), i.e. some QoS supports. This section will study how SVC can be used for such kind of network with the example of IEEE 802.11e wireless local area networks (WLANs). First, a brief background on 802.11e WLANs is given. Then a cross layer design is proposed for video streaming to fully utilize the benefits of SVC and QoS-enabled WLANs.

5.2.1 Review on IEEE 802.11e WLANs

Given the growing popularity of real-time services and multimedia based applications, it has recently become more critical to tailor IEEE 802.11 medium access control (MAC) protocol to meet the stringent requirements of such services. The IEEE 802.11 working group has developed a new standard known as the IEEE 802.11e [51] to provide QoS support. IEEE 802.11e defines a single coordination function, called the hybrid coordination function (HCF), which includes two medium access mechanisms: contention-based channel access and controlled channel access. The contention-based channel access is referred to as enhanced distributed channel access (EDCA) which extends the legacy distributed coordination function (DCF)[52] by providing the MAC layer with per-class service differentiation. In particular, EDCA defines four queues in one network node, and each queue is mapped onto one access category (AC). Each AC is designed to carry particular prioritized traffic as illustrated in Figure 5.4. $AC(n)$ has higher priority (i.e. more protected) than $AC(n-1)$. Each AC is assigned a certain contention window (CW) and arbitrary inter-frame space (AIFS) parameters to determine its priority ².

5.2.2 Optimized Cross Layer Design

With the per-class differentiation for video packets, packets with lower importance could be discarded first to protect more important packets. For example, in [53], for MPEG-4 FGS [14] stream, the base layer and enhancement layer packets are transmitted over the IEEE 802.11e WLAN with different ACs. The received video quality improved when the traffic is overloaded. This could be considered as a kind of **QoS mapping** mechanism, where video packets are assigned with different priorities and mapped to different ACs

²In general, higher values of CW and AIFS are assigned to a lower priority queue as these parameters lead to a less aggressive access to the common wireless channel. Each AC performs the back-off procedure individually according to its CW and AIFS parameters

Figure 5.4: The four access categories in an EDCA node

according to their priorities. This is also a kind of scheme for unequal error/loss protection (UEP/ULP) to explore the fine granularity scalability provided by SVC. Such a ULP idea has been implemented differently in different layers. For example, in the application layer or the MAC layer, the ULP is often provided using different FEC code [47, 48, 49] or different ARQ strategies [50]. In the network layer, the DiffServ is often used to provide ULP [54]. There are also quite a few physical layer approaches such as unequal distribution transmission power [55] and using differential modulations. Most of these existing ULP approaches only consider a single end-to-end connection and focus on optimally distributing network resource among different priorities under the constraint of a fixed total network resource. However, from the entire network point of view, the resource distribution in one connection is not independent of other competing connections. In other words, the ULP adjustment at one user will affect other competing users. Thus, the ULP strategy for one user should aim at not only maximizing its own video quality but also minimizing the bad effort to other users. To achieve this objective, it is necessary to

CHAPTER 5. APPLICATION OF 3D WAVELET BASED SVC FOR VIDEO STREAMING

consider both the video streaming characteristics and network behavior, i.e. cross-layer design is required. In this section, a practical cross-layer design is proposed to achieve optimized streaming over IEEE 802.11e WLANs with SVC.

As the SVC and IEEE 802.11e's EDCA are designed with no knowledge of each other, hence the marriage of the two requires additional entities to provide cross-layer cooperation. Targeting at achieving optimal video quality delivery using SVC over EDCA, the cross-layer design proposed here features the so called *soft capacity* property, which adaptively sacrifices video quality of each existing user to accommodate new comers to the network. In the proposed design, the application layer and network layer cooperate to decide the optimal transmission strategy at each layer. In brief, based on the detected network available bandwidth at the network layer, the application layer decides the encoding strategy and produces video streams that fit into the available bandwidth with the best possible video quality (i.e. macro rate control). However, the slower timescales at the application layer makes it difficult to respond to rapid bandwidth variation. To cope with this bandwidth variation at a smaller time, especially a sudden downside change in the available bandwidth, using the traffic information passed down from the application layer (i.e. packet priority), the network layer reacts with a further transmission rate adjustment subjected to minimum distortion. In particular, at the network layer, through an adaptive QoS mapping mechanism, micro rate control is applied to enforces ULP that preemptively sacrifices video packet with low importance to protect the transmission of those with high importance. This combination of rate controls, where the application layer and network layer perform *macro* and *micro* rate controls respectively, provides all timescale rate adaptation for video delivery over EDCA and hence ensures optimization of video streaming over the IEEE 802.11e WLAN.

Figure 5.5 presents the block diagram of the proposed design, it could be seen that the application layer consists of two components, namely optimal bit allocation and SVC

Figure 5.5: Cross layer design for video streaming over IEEE 802.11e WLAN

packetization with relative priority index (RPI). The macro rate control is achieved by the coordination between the bandwidth estimation and optimal bit allocation components, where bandwidth estimation component measures the availability of the network capacity and indicates it to the optimal bit allocation component for distortion minimization. The micro rate control is achieved by the coordination between the SVC packetization with RPI and adaptive *QoS mapping* components, where each video packet is marked with an RPI indicating its importance, then the adaptive QoS mapping adaptively directs the packets to the appropriate EDCA prioritized queues for transmission so as to maximize the transmission of important video packets. Note that here only AC2 and AC1 are used here to transport video data because AC3 is supposed for voice service[51].

Although there is a number of cross-layer studies for multimedia transmission over QoS enabled networks [54, 56, 57, 58, 59], many of these designs either employ application-centric or MAC centric approach. In contrast, the design proposed here uses an integrated approach where coordination crosses both of the application layer and the network layer. In the following subsections, each of the components in the proposed design will be

explained in details.

5.2.3 Application Layer: Macro Rate Control and Priority Assignment

Macro rate control at application layer is essentially the optimum bit allocation for the 3D wavelet based SVC. It truncates the original compressed bitstream (video source) according to the estimated bandwidth from the network layer as illustrated in Section 3.3. Note that the optimum bit allocation here is the same as the "Bit Allocation" component in the developed end-to-end streaming system for the best effort Internet (see Section 5.1.2), which introduces a time interval $T_{interval}$ and a leaky bucket to generate a CBR-like bit stream for the bottom layer. The difference is that the bandwidth information R here is given by the network layer instead of from application layer itself.

In order to provide adaptive QoS mapping in the network layer (i.e. transport different packets with different ACs), it requires the application layer to provide the relative priorities of video packets. In this research, the concept of relative priority index proposed in [54] is applied to categorize different video packets. In particular, the authors in [54] focused on a specific type of RPI for the H.263 codec, relative loss index (RLI), and calculated the RLI by summarizing the normalized "video factors" (e.g. number of Intra-coded MB, motion vector magnitude, etc) with appropriate weighting as

$$RLI_i = \sum_{n=1}^{N_{VF}} W^n \times \frac{VF_i^n}{m_i^n} \quad (\text{Eq. 5.4})$$

where N_{VF} stands for the number of video factors to be considered, W^n for the corresponding weight factors, VF_i^n for the magnitude of video factor VF activity n , and m_i^n for sampling mean of VF^n for the i -th packet. The problem of such a general RLI model is that it can hardly approximate the actual loss impact in MSE. In [56], the authors extended their method to provide a more systematic RPI association by proposing a

complicated MB-based corruption model. However, even with the new model, it is still not easy to estimate the loss impact of one packet to the MSE of the received video due to the drifting problem caused by the close-loop prediction in hybrid video codecs.

It is relatively easy to approximate the packet loss impact for the developed SVC. This is because the open-loop structure of MCTF significantly reduces the effects of the drifting problem. In particular, the loss impact of a packet is defined as the corresponding distortion increase in reconstructed video in the case that the packet is lost while all other packets are correctly received. Mathematically, the loss impact of the i -th packet in the T-band B of the color component C is calculated as

$$W_C \times G_B \times (D_{B,i-1} - D_{B,i}), \quad (\text{Eq. 5.5})$$

where W_C is the weight factor ($W_C = 1$ by default) for the color component C , G_B is the energy gain associated for T-band B and $D_{B,i}$ denotes the distortion of T-band B decoded with data from packet 0 to i .

Moreover, in order to enable smooth adjustment of the QoS mapping in the network sublayer, the loss impact values obtained in Eq. Eq. 5.5 are mapped into integer RPI values (e.g. 0-63 with 8 bits representation) which are then uniformly distributed into different packets. The detailed procedure of generating RPI values is summarized as follows. First, all the packets are sorted according to their calculated loss impact values in a decreasing order. Then, for each packet i , its position Pos_i in the sorted list is identified. Assuming the RPI values ranging from 0 to $M - 1$ and the packets with smaller RPI value are more important, for a total number of N packets, RPI can be calculated as

$$RPI_i = \text{floor}(Pos_i / N \times (M - 1)). \quad (\text{Eq. 5.6})$$

Figure 5.6 shows an example of the loss impact results in MSE for different packets³. From the figure, it can be seen that some packets have extremely large loss impact. These packets are the first packets of each T-band stream, and their loss will cause the entire T-band undecodable. Figure 5.7 shows the corresponding RPI for each packet with 8-bit quantization, which illustrates the uniform distribution of RPI values.

Figure 5.6: An example of the loss impact results

Figure 5.7: An example of the RPI value for each packet

5.2.4 Network Layer: Bandwidth Estimation and Adaptive QoS Mapping

Bandwidth estimation is an important component in the proposed cross-layer design. The application layer and network layer depend on this estimation to achieve macro rate control and optimize QoS. Its main role is to passively measure the network condition

³Sequence: Stefan CIF, with two levels of MCTF, bit rate 600kbps, packet size 500 bytes, $T_{interval} = 32/30\text{sec}$

during a macro interval $T_{interval}$ (same with the time interval used for the Optimum bit allocation at the application layer) and provide estimation of available capacity that a node can access in the next macro interval. There are several methods proposed in the literature, from earlier work of Cprobe [60], to some popular proposals of Spruce [61], Pathload [62], to some recent development of Probegap [63], wireless traffic probe [64] and IdleGap [65]. The "Feedback Control Scheme" introduced at Section 5.1.2 is also a kind of bandwidth estimation method. While many of these approaches deal with bandwidth estimation of a communication path in a wide area network (WAN), wireless traffic probe [64] and IdleGap [65] are specifically designed for the IEEE 802.11e DCF MAC protocol.

The IdleGap approach in [65] is adopted here to estimate available bandwidth $\tilde{R}$ during a macro interval $T_{interval}$. Since bandwidth will be shared among all nodes, [66] introduced a simple method to allow each node to utilize its share of the available bandwidth $\tilde{R}$. In particular, during each $T_{interval}$, each node not only estimates $\tilde{R}$, using IdleGap, but also detects and counts the number of different nodes, $\tilde{N}$, that transmit AC2 packets. This is because the detection of a video packet from a node indicates the participation of multimedia transmissions since the multimedia transmissions are usually in the form of continuous streams. If a node also transmits $\tilde{T}$ video data during the current macro interval⁴, then by the end of the current interval, the node could estimate its available bandwidth R for the next interval simply by [66]

$$R = \tilde{T} + \frac{\tilde{R}}{\tilde{N} + 1}, \quad (\text{Eq. 5.7})$$

which is then passed to the application layer to guide the optimum bit allocation to control the rate of the video stream (i.e. macro rate control).

⁴Note that, unlike UDP in the Internet, each successfully transmitted packet in WLAN is acknowledged by receiving end, therefore calculation of $\tilde{T}$ is straightforward.

CHAPTER 5. APPLICATION OF 3D WAVELET BASED SVC FOR VIDEO STREAMING

Another important task at the network layer is adaptive QoS mapping which provides the essential ULP function to increase protection for more important packets at the expense of losing less important packets. It refers to the mapping from the priority of the SVC traffic (i.e. RPI values) to the priority of the two considered EDCA queues, AC2 and AC1. Through the level of inequality the amount of sacrifices of the low priority traffic group in terms of packet drops can be induced, which in turn controls the total volume of traffic transmitted on the channel. This mapping ensure that dropped packets are most likely to happen at AC1 and hence achieves micro rate control in an indirect manner during a macro interval.

The studies of IEEE 802.11e EDCA characteristics in [66] revealed the service rate ratio of AC2 and AC1 and proposed the service ratio for the mapping to be set to two for ULP. As result, $2/3$ of the top prioritized SVC traffic needs to be transmitted via AC2 and the remaining via AC1. In the macro rate control, the application layer controls the source rate based on the bandwidth R estimated by the network layer within an macro interval. Furthermore, the RPI values of video packets is uniformed distributed. With this advantage of knowing the volume characteristics of generated video traffic, it is hence easy for the adaptive QoS mapping scheme to maintain a two to one service ratio of AC2 to AC1 for the mapping based on the remaining traffic volume rather than the fixed total traffic volume. The QoS mapping algorithm, according to [66], is detailed in **Algorithm 2**.

Within a macro interval packet drops are inevitable due to the variation in bandwidth usage of the IEEE 802.11e WLAN. the mapping point hence will float and change based on the situation of the network condition. If packet dropping happens, especially under extreme conditions where available bandwidth plunges dramatically, some level of packet dropping may also occur at AC2 which may result in some significant drop in the video

quality. It is therefore necessary to adjust the level of loss protection using packet dropping event as an indicator, e.g. decrease the value of mapping point upon a detection of packet drops so that lesser packets will be inserted to AC2 and hence loss protection at AC2 is strengthened. This is exactly what **Algorithm 2** will perform within a macro interval.

Algorithm 2 Adaptive QoS Mapping Algorithm at Network Layer

```

// RPI ranges from 0 to  $M - 1$ .
//  $v_i$ : remaining traffic volume of each  $RPI = i$  (in bits).
//  $c_i$ : expected traffic volume of each  $RPI = i$  during a macro interval.
//  $\Lambda$ : Mapping point, packets with  $RPI < \Lambda$  is inserted to AC2, otherwise to AC1.

for Each macro interval  $T_{interval}$  do
  for each  $i$  in  $[0, M-1]$  do
     $v_i = c_i$ 
  end for
   $\Lambda = \max(p)$  such that  $\sum_{0 \leq i \leq p} v_i = \frac{2}{3} \sum_{0 \leq i \leq M-1} v_i$ 
  for Each packet (with  $RPI = i$  and size  $s$  in bits) departing the network
    (either successfully transmitted or lost in the network) do
       $v_i = v_i - s$ 
       $\Lambda = \max(p)$  such that  $\sum_{0 \leq i \leq p} v_i = \frac{2}{3} \sum_{0 \leq i \leq M-1} v_i$ 
    end for
  end for
end for

```

5.2.5 Experimental Results

To test the performance of the cross-layer design for streaming SVC over the IEEE 802.11e WLAN, some experiment is done. In the experiment, the first 256 frames of the "Table-tennis" CIF sequence encoded 3D wavelet SVC with two levels of MCTF, are considered. The sequence is repeatedly transmitted for simulation time requirement. The maximum video packet size is set to 500 bytes. The RPI ranges between 0 and 63. The macro interval $T_{interval}$ is set to 32/30 seconds.

The experiment is conducted based on the network simulator NS-2 [46]. The considered network is a single hop private infrastructure WLAN. Figure 5.8(a) shows the

CHAPTER 5. APPLICATION OF 3D WAVELET BASED SVC FOR VIDEO STREAMING

experiment setup. In particular, the experiment starts with a single node, node_0. For every four macro interval duration, a new node is added to the WLAN, until the number of nodes reaches 15. The macro rate control ranges from a minimum of 500kbps to a maximum of 1 Mbps. When a node is added to the network, it shall start transport the video sequence to the base station with an initial rate of 1Mbps. A comparison is made with the performance results from that of DCF and EDCA (AC2). The former refers to the legacy DCF following the IEEE 802.11b standard [52]). The later refers to the EDCA using AC2 for video traffic with no cross-layer consideration. Y component PSNR is taken here as the video quality measurement.

Figure 5.8: WLAN experiment setup and PSNR for different streaming settings

Figure 5.8(c-d) plot the PSNR performance measured for video received from node_0 under DCF and EDCA (AC2). Two fixed source rate, 500kbps and 1Mbps are used. As can be seen, DCF supports merely three (seven) nodes for the case of 1Mbps (500kbps)

of source rate before the PSNR plunges quickly to un-acceptable quality (less than 28dB) due to network congestion. A simple calculation reveals that the maximum achievable network throughput is no more than 4 Mbps. Similar observation is made for EDCA (AC2). This is because when source rate is left uncontrolled, the network may be pushed to saturation where the protocol operates at a lower throughput level.

For the cross-layer design with the macro rate control in place, each node varies its source rate from no less than 500kbps to no more than 1Mbps according to the estimated bandwidth. Its PSNR result in Figure 5.8(b) shows clearly the performance advantage. The *soft capacity* property is also illustrated. When the network consists of a small number of nodes, each node receives high quality video. As the number of nodes increases, all nodes adapt themselves to the changed network condition to accommodate more nodes by reducing sending rate. Even with the minimum macro rate to 500kbps, the micro rate control could still preemptively drops video packet of low importance via AC1 so that quality of video streams could be maintained.

Figure 5.9 compares the packet loss ratio (in percentage). As showed in Figure 5.9(c-d), DCF and EDCA(AC2) share the similar behavior in loss ratio. Without QoS mapping, important video packets maybe be dropped and quality of video cannot be maintained when packet loss occurs. On the other hand, the ULP with the adaptive QoS mapping protects packets on AC2 by sacrificing packets on AC1. There is hardly any loss in AC2 as showed in Figure 5.9(a). Figure 5.9(b) further plots the packet loss ratio from the application's point of view. For clarity of presentation, the 64 RPI values are partitioned into three groups. As can be seen, groups of lower importance suffers high packet loss compared with that of higher importance.

Figure 5.10 shows the action of macro rate control. As more nodes are added to the WLAN, the detected available bandwidth drops and hence the source rate at node_0 slowly decreases from the initial 1Mbps to the eventual minimum rate of 500kbps.

CHAPTER 5. APPLICATION OF 3D WAVELET BASED SVC FOR VIDEO STREAMING

Figure 5.9: Packet loss ratio for different streaming settings on IEEE 802.11e WLAN

Figure 5.10: Illustration of macro rate control in the cross-layer design

5.3 Summary

In this section, two different streaming systems are described to illustrate the potential application of the developed 3D wavelet based scalable video codec. For the best-effort Internet without any QoS support, an end-to-end system is designed. Its main component is the end-to-end feedback control scheme at the application layer to estimate bandwidth. For QoS-enabled networks, with the example of IEEE802.11e WLAN, a cross-layer design is proposed to optimize video streaming sessions. The main components include (1) the SVC packetization with RPI that marks the importance of video packets and macro rate control at the application layer and (2) bandwidth estimation and ULP with adaptive QoS mapping at the network layer. In both systems, the scalability property of SVC are explored. Although only the quality scalability is illustrated here. The spatial and temporal scalability can be easily exploited by introducing additional QoS parameters (e.g. frame rate, spatial resolution etc) to guide the bit allocation component in the two streaming systems as showed in Figure 5.5.

Chapter 6

Conclusions and Future Work

Scalable video coding, which aims to provide spatial, temporal, and quality scalability at bit stream level, becomes more and more attractive. In this thesis, a simple scalable video coding scheme based on the integration of motion compensated temporal filtering (MCTF) and JPEG2000 has been proposed. In particular, MCTF is used to decompose video frames along the temporal axis into temporal bands. Multiple levels of MCTF enable powerful exploration of multiple frames redundancies which is hardly achievable by conventional hybrid video codec. The generated T-bands are then further encoded based JPEG2000. By exploring the unique features of MCTF and JPEG2000, the proposed simple codec not only achieves a competitive RD performance but also has the properties of high scalability and error resilience. The main contributions of the developed codec lie in its simple architecture and the optimum bit truncation across the T-bands based on the quality layer concept in JPEG2000. Although several codecs based on similar integrations have been proposed in the literature [28, 31, 30, 32], most of them have complex coding structure in order to pursue superior R-D performance against the state-of-art H.264/AVC. More over, with the proposed simple codec, many features in JPEG2000, such as ROI coding and error resilience, can be easily utilized. Furthermore, the proposed simple codec also serves as a good framework for the study of other important issues related to 3D wavelet based coding, such as the problem of constant quality (Chapter 4)

CHAPTER 6. CONCLUSIONS AND FUTURE WORK

and the transportation of video stream over networks (Chapter 5).

Due to the inaccuracy in the theoretical distortion analysis for MCTF (see Section 3.3.1) and the fact that the Lagrange-multiplier based optimum allocation method only aims to minimize average distortion instead of quality difference among reconstructed frames, the fluctuation of quality among the reconstructed frames is usually very severe after the optimum rate allocation (see Section 4.1). To address this problem, a constant quality aimed bit allocation algorithm has been proposed to reduce this quality fluctuation, which consist of two steps. First, by analyzing the distortion relations between T-bands and reconstructed frames, the energy gain for T-bands of different levels is adjusted so that different classes of reconstructed frames have a close distortion. Second, after the optimum rate allocation with the new energy gains, a smoothing method is applied to adjust T-band rate so that the MSE distribution for the T-bands at the same level are is smoothed. Experimental results have showed the proposed method could effectively reduce the PSNR fluctuation problem.

Scalable video coding schemes are intended to encode the video signal once at highest resolution and quality, but enable decoding from partial streams depending on the specific bit rate and resolution required by a certain application. This enabled a simple and flexible approach for transmission over heterogeneous networks, additionally providing adaptability for bandwidth variations and error conditions. The thesis investigates some potential usage of the developed 3D wavelet based scalable video codec with example of stored video streaming over different computer networks. In particular, two different streaming systems are developed. For the best-effort Internet, where there is no any QoS support, an end-to-end system is designed. An end-to-end feedback control scheme is performed at the application layer to estimate the available network bandwidth, which is used to determine the video stream bit rate for the streaming session. For QoS-enabled networks, with the example of IEEE802.11e WLAN, a cross-layer design is proposed to

optimize received video quality. At the application layer, the SVC stream is packetized with RPI that marks the importance of video packets. At the network layer, available bandwidth is estimated and passed to the application layer to guide the truncation of video bit stream. In addition, adaptive QoS mapping is performed based on the RPI value of each packets in order to protect packets with high priority. In both systems, the scalability property of SVC is explored.

6.1 Future Work

The list of future work which can be carried is almost limitless. There are many ways to further extend the developed scalable video codec with greater performance and or utilize it for more services/applications. Some of them are listed as follows:

- (i) Further optimization in terms of RD performance. There are many ways to improve the existing work. For MCTF based coding, it is often found that the performance of this type of codecs, including the proposed codec, deteriorates at lower rates. This is mainly due to the fact that motion vectors as optimized for higher rates consume a big percentage of the overall rate, unless the motion information is encoded in a scalable fashion as well. Possible approaches for scalable motion encoding include scaling of accuracy (e.g., from quarter-pixel precision to half-pixel), of the spatial resolution (in terms of length of vectors depending on image size, and block size of motion vectors), and temporal resolution of motion vector fields. To use these scalable motion vector coding, effects of different motion vectors usage during MCTF analysis and synthesis needs to be minimized, which means that scalable motion encoding should be inherently tied with the variation of video signal resolution.
- (ii) Extends the codec for multi-view video coding. Multi-view video [67] is captured at different viewing angles at the same time and hence has one more dimension than

CHAPTER 6. CONCLUSIONS AND FUTURE WORK

normal video data. In addition to the temporal and spatial redundancy within it, the view redundancy within view adjacent frames also needs to be exploited. To extend the developed codec for this kind of video, one more dimension of wavelet transform along the view axis is required. Further investigations on how the transform is executed and how to code any additional motion vectors are then needed.

- (iii) Transporting scalable video over networks. The two streaming systems described in this thesis illustrate the potential usage of the developed 3D wavelet based scalable video codec. There are still many applicable improvements for the two developed systems. For example, how to apply error concealment with special attention to the unique property of 3D wavelet and bit-plane coding which are basis for the developed codec? Because all existing error concealment schemes are developed for conventional hybrid video codec, which are quite different from the 3D wavelet based SVC. For the end-to-end streaming in the Internet, some special unequal error protection schemes at the application layer could also be developed to further optimize video quality received. For networks with different kinds of QoS support from that of IEEE 802.11e WLAN, different designs for the streaming system might be necessary to optimize received video quality.

References

- [1] Mohammed Ghanbari, *Video Coding - An introduction to standard codecs*, The Institution of Electrical Engineers, 1999.
- [2] J-R Ohm, "Advances in scalable video coding," *Proc. of the IEEE*, vol. 93, no. 1, pp. 42–56, Jan. 2005.
- [3] A. Secker and D Taubman, "Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting," *Proc. IEEE Int. Conf. Image Processing*, vol. 2, 2001.
- [4] L. Luo, J. Li, S. Li, Z. Zhuang, and Y. Q. Zhang, "Motion compensated lifting wavelet and its application in video coding," *Int. Conf. Multimedia & Expo (ICME)*, pp. 365–368, 2001.
- [5] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," *Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1793–1796, 2001.
- [6] "ISO/IEC 15444-1: Information technology - JPEG2000 image coding system - part 1: Core coding system," 2000.
- [7] "Video codec for audiovisual services at px64kbit/s," *ITU-T Recommendation H.261 version 1*, 1990.
- [8] "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbit/s - part2:video," *International standard ISO/IEC 1117-2*, Nov. 1991.

REFERENCES

- [9] "Coding of audio-visual objects - part2: visual," *ISO/IEC 14496-2 (MPEG-4 visual version 1)*, Apr. 1999.
- [10] "Video coding for low bit rate communication," *Draft ITU-T Recommendation H.263*, Jul. 1995.
- [11] "Generic coding of movign pictures and associated audio information- part2: video," *International standard ISO/IEC 12818-2*, Nov. 1994.
- [12] "Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264/ISO/IEC 14496-10 AVC)," *Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050*, 2003.
- [13] Weiping Li, "Overview of fine granularity scalability in MPEG-4 video standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 301–317, Mar. 2001.
- [14] "Coding of audio-visual objects - part2: visual, amendent 4: streaming video profile," *ISO/IEC 14496-2/FPDAM4*, Jul. 2000.
- [15] M. van der Schaar and H. Radha, "Hybrid temporal-SNR fine-granular scalability for internet video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 318–331, Mar. 2001.
- [16] R. Yan, F. Wu, S. Li, R. Tao, and Y. Wang, "Efficient video coding with hybrid spatial and fine-grain SNR scalabiities," *Proc. Visual Comm. and Image Processing (VCIP)*, vol. 4671, pp. 850–858, 2002.
- [17] F. Wu, S. Li, and Y-Q. Zhang, "A framework for efficient progressive fine granularity scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 332–344, Mar. 2001.
- [18] X. Sun, F. Wu, S. Li, W. Gao, and Y-Q. Zhang, "Macroblock-based progressive fine granularity scalable video coding," *Proc. IEEE Inter. Conf. Multimedia and Expo (ICME)*, 2001.

REFERENCES

- [19] H-C. Huang, C-N. Wang, and T. Chiang, "A robust fine granularity scalability using trellis-based predictive leak," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 372–385, Jun. 2002.
- [20] R. E. Crochiere, S. A. Weber, and J. L. Flanagan, "Digital coding of speech in sub-bands," *Bell System Technical Journal*, vol. 55, pp. 1069–1085, 1967.
- [21] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*, chapter 4: Image Transform, pp. 143–208, Boston/Dordrecht/London: Kluwer Academic Publishers, 2002.
- [22] J. M. Shapiro, "Embedded image coding using zero-trees of wavelet coefficients," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3445–3462, Dec. 1993.
- [23] A. Said and W. Pearlman, "A new, fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 3, pp. 243–250, Jun. 1996.
- [24] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Trans. Image Process.*, vol. 9, no. 6, pp. 243–250, Jun. 1996.
- [25] G. Karlsson and M. Vetterli, "Subband coding of video signals for packet switched networks," *Proc. SPIE, Visual Communi. Image Process.(VCIP)*, vol. 845, pp. 446–456, 1987.
- [26] C. I. Podilchuk, N.S. Jayant, and N. Farvardin, "Three-dimensional subband coding of video," *IEEE Trans. Image Process.*, vol. 4, no. 2, pp. 125–139, Feb. 1995.
- [27] J-R Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 559–571, Sep. 1994.
- [28] Choi S-J and J. W. Woods, "Motion-compensated 3D subband coding of video," *IEEE Trans. Image Processing*, vol. 3, no. 2, pp. 155–167, Feb. 1999.

REFERENCES

-
- [29] R. Xiong, J. Xu, F. Wu, S. Li, and Y.-Q. Zhang, "Optimal subband rate allocation for spatial scalability in 3D wavelet video coding with motion aligned temporal filtering," *Proc. of SPIE, Visual Commun. Image Processing (VCIP)*, vol. 5960, pp. 381–392, 2005.
 - [30] R. Schafer, H. Schwarz, D. Marpe, T. Schierl, and T. Wiegand, "MCTF and scalability extension of H.264/AVC and its application to video transmission, storage, and surveillance," *Proc. of SPIE, Visual Comm. and Image Process (VCIP)*, vol. 5960, 2005.
 - [31] K. Hanke, J. R. Ohm, and T. Ruster, "Adaptation of filters and quantization in spatio-temporal wavelet coding with motion compensation," *Proc. of the Picture Coding Symposium*, pp. 49–54, Apr. 2003.
 - [32] J. Xu, Z. Xiong, S. Li, and Y.-Q. Zhang, "3-D embedded subband coding with optimal truncation (3-D ESCOT)," *J. Applied and Computational Harmonic Analysis: Special Issue on Wavelet Applications in Engineering*, vol. 10, pp. 290–315, May 2001.
 - [33] X. M. Zhang, A. Vetro, Y. Q. Shi, and H. Sun, "Constant quality constrained rate allocation for FGS-coded video," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 121–130, Feb. 2003.
 - [34] J. Cai, Z. He, and C. W. Chen, "A novel frame-level bit allocation based on two-pass video encoding for low bit rate video streaming applications," *Journal of Visual Communications and Image Representation*, vol. 17, pp. 783–798, 2003.
 - [35] P. Chen and J. W. Woods, "Bidirectional MC-EZBC with lifting implementation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 10, pp. 1183–1194, Oct. 2004.
 - [36] T. Ruster, K. Hanke, and J. R. Ohm, "Transition filtering and optimized quantization in interframe wavelet video coding," *Proc. of SPIE, Visual Comm. and Image Processing (VCIP)*, vol. 5150, pp. 682–393, 2003.

REFERENCES

-
- [37] A. Mavlankar, S. E Han, C. L. Chang, and B. Girod, "A new update step for reduction of PSNR fluctuations in motion-compensated lifted wavelet video coding," *Proc. IEEE Inter. Workshop Multimedia Signal Processing (MMSP)*, Nov. 2005.
 - [38] D. Wu, Y. T. Hou, W. Zhu, Y. Q Zhang, and J. M. Peha, "Streaming video over the internet: Approaches and directions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 282–300, Mar. 2001.
 - [39] D. Wu, Y. T. Hou, W. Zhu, H. J. Lee, T. Chiang, Y. Q. Zhang, and H. J. Chao, "On end-to-end architecture for transporting MPEG-4 video over the internet," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 6, pp. 923–941, Sep. 2000.
 - [40] D. Wu, Y. T. Hou, B. Li, W. Zhu, Q. Zhang Y, and H. J. Chao, "An end-to-end approach for optimal mode selection in internet video communication: Theory and application," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 6, Jun. 2000.
 - [41] D.M. Chiu and R. Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks," *Comp. Networks and ISDN Systems*, vol. 17, no. 1, 1-14 1989.
 - [42] S. Floyd and K. Fall, "Promoting the use of end-to-end congestion control in the internet," *IEEE Trans. Networking*, vol. 7, no. 4, pp. 458–472, Aug. 1999.
 - [43] Z. G. Li, C. Zhu, N. Ling, X. K. Yang, G. N. Feng, S. Wu, and F. Pan, "A unified architecture for real-time video-coding systems," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 6, pp. 472–486, Jun. 2003.
 - [44] Z. Ni, Z. Chen, and K.N. Ngan, "A real-time video transport system for the best-effort internet," *Signal Processing: Image Communication*, vol. 20, no. 3, pp. 277–293, 2005.
 - [45] H. Schulzrinne, S. Casner, R. Frederich, and V. Jacobson, "Rtp: A transport protocol for real-time applications," *Internet Engineering Task Force, RFC 3550*, July 2003.

REFERENCES

- [46] K. Fall and K. Varadhan (ed.), "The NS manual [online]," Available at <http://www.isi.edu/nsnam/ns/doc/index.html>.
- [47] X. K. Yang, C. Zhu, Z. Li, G. Feng, S. Wu, and N. Ling, "A degressive error protection algorithm for MPEG-4 FGS video streaming," *Proc. IEEE Int. Conf. Image Processing*, pp. 737–740, 2002.
- [48] M. van der Schaar and H. Radha, "Unequal packet loss resilience for fine-granular-scalability video," *IEEE Trans. Multimedia*, vol. 3, no. 4, pp. 381–393, Dec. 2001.
- [49] T. Stockhammer, "Progressive video transmission for packet lossy channels exploiting feedback and unequal erasure protection," *Proc. of IEEE Int. Conf. Image Processing*, 2002.
- [50] C. Liu and S. S. Chen, "Providing unequal reliability for transmitting layered videostreams over wireless networks by multi-ARQ schemes," *Proc. IEEE Int. Conf. Image Processing*, pp. 100–104, 1999.
- [51] IEEE, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment 8: Medium Access Control (MAC) Quality of Service Enhancement," IEEE std 802.11e-2005, IEEE, New York, 2005.
- [52] IEEE, "IEEE Standard for Information Technology - Telecommunications and Information Exchange between Systems - Specific Requirements Part 11: Wireless LAN MAC and PHY Specifications," IEEE std 802.11-1999, IEEE, New York, 1999.
- [53] Y. Zhang, Z. Ni, C. H. Foh, and J. Cai, "Improving scalable video transmission over IEEE802.11e networks," *Proc. of SPIE, Optics East Conf. Multimedia and Applications viii*, Oct. 2005.
- [54] J. Shin, J. W. Kim, and C. C. J. Kuo, "Quality-of-service mapping mechanism for packet video in differentiated service network," *IEEE Trans. Multimedia*, vol. 3, no. 2, pp. 219–231, Jun. 2001.
- [55] Q. Zhang, W. Zhu, Z. Ji, and Y.-Q. Zhang, "Power-minimized bit allocation for video communication over wireless channels," *Special issue on wireless video in IEEE Trans. on Circuit and System for Video Technology*, 2002.

REFERENCES

- [56] J.-G. Kim, J. Shin, J. W. Kim, and C. C. J. Kuo, "Coordinated packetlevel protection with a corruption model for robust video transmission," *Proc. Visual Communication and Image Processing (VCIP)*, pp. 410–421, 2001.
- [57] W. Kumwilaisak, Y. T. Hou, Q. Zhang, W. Zhu, C.-C. Jay Kuo, and Y.-Q. Zhang, "A cross-layer quality-of-service mapping architecture for video delivery in wireless network," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 10, Dec. 2003.
- [58] A. Ksentini, M. Naimi, and A. Gueroui, "Toward an improvement of h.264 video transmission over ieee 802.11e through a cross-layer architecture," *IEEE Communications Magazine*, pp. 107–114, Jan. 2006.
- [59] M. van der Schaar and S. Shankar N, "Cross-layer wireless multimedia transmission: Challenges, principles, and new paradigms," *IEEE Wireless Communications*, pp. 50–58, Aug. 2005.
- [60] R. L. Carter and M. E. Crovella, "Dynamic server selection using bandwidth probing in wide-area networks," *Technical Report BU-CS-96-007, Computer Science Department, Boston University*, March 1996.
- [61] J. Strauss, D. Katabi, and F. Kaashoek, "A measurement study of available bandwidth estimation tools," *Proc ACM IMC*, 2003.
- [62] M. Jain and C. Dovrolis, "Pathload: A measurement tool for end-to-end available bandwidth," *Proc Passive and Active Measurements*, March 2002.
- [63] K. Lakshminarayanan, V. N. Padmanabhan, and J. Padhye, "Bandwidth estimation in broadband access networks," *Proc ACM IMC*, 2004.
- [64] M. Davis, "A wireless traffic probe for radio resource management and qos provisioning in ieee 802.11 wlans," *Proc. ACM International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, 2004.
- [65] H. K. Lee, V. Hall, K. H. Yum, K. I. Kim, and E. J. Kim, "Bandwidth estimation in wireless lans for multimedia streaming services," *Proc. IEEE International Conference on Multimedia and Expo*, 2006.

REFERENCES

- [66] C. H. Foh, Y. Zhang, Z. Ni, J. Cai, and K. N. Ngan, "Optimized cross-layer design for scalable video transmission over the ieee 802.11e networks," *to appear in IEEE Trans. Circuits and Systems for Video Technology.*, 2007.
- [67] ISO/IEC JTC1/SC29/WG11 N6501, "Requirements on multi-view video coding," Redmond, USA, July 2004.

Related Publications

- Y. Zhang, Z. Ni, C. H. Foh, and J. Cai, Improving scalable video transmission over IEEE 802.11e networks, *Proc. of SPIE*, Vol. 6015, Oct. 2005.
- Z. Ni and J. Cai, Constant quality aimed bit allocation for 3D scalable video coding, *Proc. 2006 IEEE International Conference on Multimedia & Expo (ICME 2006)*, July, 2006.
- C. H. Foh, Y. Zhang, Z. Ni, J. Cai, and K. N. Ngan, Optimized Cross-Layer Design for Scalable Video Transmission over the IEEE 802.11e Networks, *accepted by IEEE Trans. Circuits and Systems for Video Technology*..
- Y. Zhang, Z. Ni, C.H. Foh and J. Cai, Retry Limit Based ULP for Scalable Video Transmission over the IEEE 802.11e WLANs, *IEEE Communications Letters*, Vol. 11, No. 6, pp498-500, June 2007.