Bit allocation techniques for video coding and broadcasting applications

Yao, Wei

2013

Yao, W. (2013). Bit allocation techniques for video coding and broadcasting applications.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/54879>

<https://doi.org/10.32657/10356/54879>

**Bit Allocation Techniques
for
Video Coding and Broadcasting Applications**

YAO WEI

School of Electrical & Electronic Engineering

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

2013

Acknowledgements

First and foremost, I would like to express my deep and sincere gratitude to my supervisors, Prof. Lap-Pui Chau and Prof. Susanto Rahardja, for their thoughtful guidance, invaluable assistance and constant encouragement throughout the whole process. I highly appreciate their kind patience and instant responses to all my queries in their busy schedules. Their wide knowledge and enthusiasm for the research have had a remarkable influence on my entire graduate study.

Special thanks to my colleague, Dr. Zhengguo Li for his professional guidance, insightful comments and helpful suggestions on all the work I have done. I also want to thank my colleagues, Dr. Haiyan Shu and Dr. Yu Wang, for their warm encouragement and friendship, which have been invaluable on both of my academic and personal life. I wish to thank all the researchers at Institute for Infocomm Research whom I've worked with. It is a great pleasure to work and study with all of them.

Last but not least, I would like to thank my family for their love and endless support. They have taught me how to deal with problems and how to adapt myself to the challenges of life. Without their support, this work would never have been accomplished.

Summary

In today's world, it is no longer uncommon to see video applications that employ a variety of transmission and storage systems, thanks to the significant progress in video coding technologies together with great improvements on network infrastructure and the increasing growth of storage capacity and computing power. With an increasing demand for digital video applications, the expectation on the quality of these applications is getting higher. Bit allocation is a classical problem in video coding and it plays an essential role in achieving better coding efficiency and better visual quality. Usually bit allocation can be considered as the constrained optimization problem, in which a given bit rate budget is distributed among different units in different application scenarios. In this thesis, the bit allocation problem for various video applications is studied. The focus of the research is on algorithm design and coding structure optimization. The goal is to provide an efficient way to utilize the available resources so as to produce optimum visual quality of the reconstructed video with the given resource constraint.

Bit allocation within a single video source is more commonly referred to as rate control. If there is a limited memory buffer at the encoder, rate control of an encoded video will reduce buffer overflow. Rate control also ensures that the output bit rate of video stream remains within the capacity of transmission channel. When conventional rate control algorithms perform bit allocation at group of pictures (GOP) level, a fixed GOP size is usually adopted without considering the correlation between frames. A new rate control algorithm in which the sizes of GOPs are adapted to the video content is proposed. Temporal redundancies among frames can be better removed and it transforms to improvements on coding efficiency.

Scalable video coding (SVC) provides a scalable scheme for "one-source-multiple-use" in heterogeneous network. The layered structure of SVC allows one single bit stream to serve multiple users with different decoding and display capabilities and/or network channels with different bandwidths. Both cross layer bit allocation for individual user and bit allocation for multiple users are studied in this work. A combined SNR and spatial scalability scheme is

proposed to optimize a specific layer according to customer composition. A motion refinement scheme is then developed to further improve the coding efficiency of that specific layer through a better bit allocation between motion and residual information. When bit allocation is unbalanced in current coding structure, an improved inter-layer prediction scheme is proposed for layers with both spatial and quality scalability and a new concept of auxiliary layer is introduced to further improve the coding efficiency.

In broadcast applications, multiple video sources are encoded in parallel and transmitted simultaneously over a bandwidth limited channel. Coding of each video independently at an equal proportion of the channel bandwidth is a straightforward method to allocate channel bandwidth but it suffers from big quality fluctuations between videos as well as within each video. A new joint bit rate allocation scheme is designed to dynamically allocate the channel bandwidth so that a more uniform video quality is achieved for various videos and the quality variation within each video is also minimized.

With recent advances in display and camera technologies, 3D video are becoming popular for next generation broadcast and streaming applications. Although there are various representation formats available for 3D video, stereoscopic seems to be the most mature technology at this moment. However, due to the increased video data compared to conventional single-view video, efficient coding format and joint bit rate allocation among multiple stereo videos on bandwidth limited transmission links are still challenging problems. A scalable coding scheme for stereo video is first designed such that regulating output bit rate of each video to match the allocated bandwidth can be achieved by simple bitstream extraction. The bit allocation problem for stereo videos is more complex as it includes both bit allocation between videos as well as between views. By exploiting human visual system, a joint bit allocation scheme is developed to do adaptive bit allocation between two views so that higher perceptual quality in each video is achieved while the variance of perceptual quality among different stereo videos is also reduced comparing with existing method.

Contents

Acknowledgements	i
Summary	iii
List of Figures	ix
List of Tables	xi
List of Symbols	xiii
Abbreviations	xv
1 Introduction	1
1.1 Background and Motivation	1
1.2 Objective and Main Contribution	5
1.3 Outline of The Thesis	7
2 Overview of Video Coding and Bit Allocation	9
2.1 A typical video compression system	9
2.2 Video Coding Standards	12
2.2.1 Earlier Codecs	12
2.2.2 H.264/AVC	13
2.2.3 Scalable Video Coding Extension of H.264/AVC	14
2.2.3.1 Temporal Scalability	17
2.2.3.2 Spatial Scalability	18
2.2.3.3 Quality Scalability	20
2.3 Rate Control for Video Coding	23
2.4 Joint Rate Allocation for Statistical Multiplexing	26
3 Adaptive GOP Rate Control Based On Keyframe Detection	29
3.1 Introduction	29
3.2 Background and Related Works	32
3.2.1 Rate Control with Fixed GOP size	32
3.2.2 Scene Change Detection Algorithm	36
3.3 Content Adaptive GOP Size Control	37
3.3.1 Video Content Variation Metric	37
3.3.2 Dynamic Threshold Model	38

3.3.3	Keyframe Detection Algorithm	44
3.4	Rate Control with Adaptive GOP size	45
3.4.1	Two Requirements on the Sizes of GOPs	45
3.4.2	Computation of $N(i)$	46
3.4.3	Bit allocation and QP determination	47
3.4.4	Algorithm flowchart	48
3.5	Experimental Results	50
3.6	Summary	56
4	Implicit Bit Allocation and Cross-Layer Bit Allocation on SVC	59
4.1	Introduction	59
4.2	Customer Oriented Combined Spatial and SNR Scalability	61
4.2.1	Motivation	61
4.2.2	A New Type of ROI and A New Performance Index	62
4.2.3	Spatial scalability of Motion Information and Residual Information	67
4.2.4	SNR Scalability of Motion Information and Residual Information	70
4.2.5	Experimental Results	70
4.2.6	Summary	74
4.3	Motion Refined Enhanced Coarse Granular Scalability	75
4.3.1	Motivation	75
4.3.2	Difference Between Motion Refined CGS and Motion Refined MGS	77
4.3.3	The Motion Refined MGS Scheme	78
4.3.4	Experimental Results	80
4.3.5	Summary	84
4.4	Balanced Inter-Layer Prediction	86
4.4.1	Motivation	86
4.4.2	Inter-Layer Prediction with Two Reference Layers	89
4.4.2.1	Experimental Results	92
4.4.3	A New Concept of Auxiliary Layers	94
4.4.3.1	Experimental Results on Auxiliary Layers	97
4.4.4	Summary	99
5	Joint Bit Allocation for Multi-Program Video Broadcast	101
5.1	Introduction	101
5.2	Joint Rate Allocation For Statistical Multiplexing in Video Broadcast Applications	104
5.2.1	Background and Related Work	104
5.2.2	System Overview	105
5.2.3	Similarity Based Complexity Analyzer	106
5.2.3.1	Existing Structure Similarity Index	107
5.2.3.2	New Complexity Analyzer	108
5.2.4	Joint Rate Allocation Scheme	110
5.2.4.1	Rate Allocation for Base Layer	110
5.2.4.2	Rate Allocation for Video Adaptation	111
5.2.5	Experimental Results	117
5.2.5.1	Experimental Setup	117
5.2.5.2	Results on Base Layer Bit Allocation	118

5.2.5.3	Results on System Bit Allocation	120
5.2.6	Summary	124
5.3	Joint Rate Allocation of Stereoscopic 3D Videos in Next-Generation Broadcast Applications	126
5.3.1	Background and Related Work	126
5.3.1.1	3D Video Representation and Coding	126
5.3.1.2	Asymmetric Stereo Video Coding and Quality assessment . . .	132
5.3.2	System Overview	134
5.3.3	Scalable Stereoscopic Video Coding	135
5.3.4	Joint Rate Allocation for Multiple Stereo Videos	139
5.3.4.1	Rate Allocation Among Videos	139
5.3.4.2	Adaptive Inter-view Rate Allocation	141
5.3.4.3	Algorithm Flowchart	143
5.3.5	Experimental Results	146
5.3.6	Summary	152
6	Conclusions and Future Works	155
6.1	Conclusions	155
6.2	Future Work	158
	Author's Publications	160
	Bibliography	163

List of Figures

2.1	Generalized block diagram of a video compression system.	10
2.2	Development of various video coding standards.	12
2.3	(a) Separate encoding/transcoding and simulcast transmission. (b) Scalable video coding and transmission	16
2.4	Overall SVC encoder structure with three spatial layers.	17
2.5	Hierarchical-B coding structure (GOP size = 8, No. of temporal levels = 4) [8]. .	18
2.6	General structuring of a scalable bitstream with a base layer and L quality enhancement layers.	21
2.7	VBR and CBR coding schemes with corresponding relations between the output bitrate and quality.	25
2.8	Procedures for a typical video rate control scheme.	26
2.9	Independent coding of multiple video signals.	27
2.10	Joint coding of multiple video signals.	27
3.1	Number of IMBs in each frame of a concatenated sequence encoded at CIF resolution and 15fps.	38
3.2	(Left) Changes on mean of IMBs (Right) Changes on standard deviation of IMBs at different resolutions	40
3.3	(Left) Changes on mean of IMBs; (Right) Changes on standard deviation of IMBs at different QPs	42
3.4	The diagram of the proposed scheme to determine GOP size	47
3.5	Flowchart of the proposed algorithm.	49
3.6	RD performance comparison on QCIF sequences	51
3.7	RD performance comparison on CIF sequences	52
3.8	RD performance comparison on 4CIF sequences	52
4.1	(a) Conventional ROI in image processing (b) New Type of ROI in SVC	63
4.2	Reference layer selection - Default	67
4.3	Reference layer selection - Case 1 and Case 2	68
4.4	RD performance for Motion Refined MGS	84
4.5	Inter-layer prediction structure by default	87
4.6	Improved Inter-layer prediction for (CIF, R_2)	87
4.7	Experimental results for Layer 2 - (CIF, R_2) when reference layer of Layer 2 is changed from Layer 1 to Layer 0	88
4.8	Experimental results for Layer 3 - (CIF, R_3) when reference layer of Layer 2 is changed from Layer 1 to Layer 0	88
4.9	The proposed inter-layer prediction structure with two reference layers	90

4.10	Comparison between proposed scheme and JSVM non-adaptive scheme for the sequence "Foreman", "Crew", "City", "Soccer".	93
4.11	The inter-layer prediction structure with Auxiliary Layer when $R_2/R_0 < 3.5$. . .	96
4.12	: Visual quality comparison between (a) existing scheme without AL and (b) the proposed scheme with AL for the sequence "Foreman" Frame #35.	99
4.13	Visual quality comparison between (a) existing scheme without AL and (b) the proposed scheme with AL for the sequence "Bus" Frame #42.	99
5.1	Structure of the proposed multiprogram video coding system	106
5.2	Illustration of the problem in attaining both inter-program fairness and intra-program smoothness for broadcasting applications.	111
5.3	R-D curve of different FGS layers.	113
5.4	Screenshot of CIF (352×288) sequences: "Foreman", "Football", "M&D" and "Harbor".	117
5.5	Screenshot of HD (1280×720) sequences: "Crew", "Bigship", "Raven" and "Shuttle_start".	118
5.6	Quality fluctuation at GOP boundaries in each video sequence at CIF resolution with the channel bandwidth = 2MB/s	123
5.7	Quality fluctuation at GOP boundaries in each video sequence at HD resolution with the channel bandwidth = 10MB/s	125
5.8	(Top) Full-resolution left and right views, and (Bottom) frame-compatible representation of stereoscopic videos.	128
5.9	Frame-compatible coding using SVC with spatial scalability.	130
5.10	Frame-compatible coding using SVC with a combined spatial and temporal scalability.	131
5.11	Basic structure of joint coding system for multiple stereo videos	136
5.12	Scalable stereo coding structure.	137
5.13	Structure of scalable stereo bitstream with L quality refinement layers for both views.	138
5.14	Flowchart for finding the optimal bit allocation by the golden section search method.	145
5.15	Screenshot of stereo videos: "Street", "Fountain", "Lake", "Rollerskate" (a) Left View (b) Right View.	146

List of Tables

3.1	Average PSNR gain for Proposed Dynamic Threshold Based Keyframe Selection	53
3.2	Processing time for Proposed Dynamic Threshold Based Keyframe Selection . .	54
3.3	Performance comparison between proposed rate control and JM rate control . .	56
4.1	The settings of five layers in experiment 4.2.5	71
4.2	Reference layer selection in Type 1	71
4.3	Experimental results in Type 1: (CIF, Low) is demanded by very few customers	72
4.4	Reference layer selection in Type 2 and 3	72
4.5	Experimental results in Type 2: (CIF, Low) is demanded by many customer but $w_{l,2,3}$ > $w_{l,2,1}$	72
4.6	Experimental results in Type 3: (CIF, Low) is demanded by many customer and $w_{l,2,3}$ < $w_{l,2,1}$	72
4.7	Major difference between motion refinement for CGS and MGS	77
4.8	QPs for the generation of motion information at each layer (Foreman)	80
4.9	PSNR (dB) of each type at two layers (Foreman)	81
4.10	QPs for the generation of MVF at each layer (Foreman)	81
4.11	PSNR(dB) value at each layer (Foreman)	81
4.12	Experimental results for MGS	83
4.13	Values of QPs for the generation of motion information at all layers	83
4.14	QP settings for each sequence	92
4.15	Bitrate increment factor for QCIF to CIF and CIF to 4CIF	95
4.16	Experimental results with $R_A = 64\text{kb/s}$	97
4.17	Experimental results with $R_A = 96\text{kb/s}$	98
5.1	Summary of different characteristics on test sequences	118
5.2	Average PSNR & variance of PSNR for base Layers of different sequences at CIF resolution (Base Layer Bandwidth = 1MB/s)	120
5.3	Average PSNR & variance of PSNR for base layers of different sequences at HD resolution (Base Layer Bandwidth = 3MB/s)	121
5.4	Average PSNR , variance of PSNR in each GOP Of different sequences and sum Of PSNR fluctuations in each sequence (CIF resolution, Channel Bandwidth = 2MB/s, GOP = 16, Framerate = 30 frame/s, Hierarchical-B structure)	122
5.5	Average PSNR , variance of PSNR in each gOP Of Different sequences and sum Of PSNR fluctuations in each sequence (HD resolution, Channel Bandwidth = 10MB/s, GOP = 30, framerate = 60 frame/s, IPP...P structure)	124
5.6	PSNR in individual view, perceived quality, and variance of perceived quality for 4 stereo videos achieved by proposed scheme (Channel bandwidth = 20MB/S)	148

5.7	PSNR in individual view, perceived quality, and variance of perceived quality for 4 stereo videos achieved by frame-compatible framework(Channel bandwidth = 20MB/S)	149
5.8	PSNR in individual view, perceived quality, and variance of perceived quality for 4 stereo videos achieved by full-resolution statistical multiplexing framework(Channel bandwidth = 20MB/S)	149
5.9	PSNR in individual view, overall perceived quality, and variance of perceived quality for 4 stereo videos achieved by proposed scheme (Channel bandwidth = 40MB/S)	150
5.10	PSNR in individual view, overall perceived quality, and variance of perceived quality for 4 stereo videos achieved by frame-compatible framework (Channel bandwidth = 40MB/S)	151
5.11	PSNR in individual view, overall perceived quality, and variance of perceived quality for 4 stereo videos achieved by full-resolution statistical multiplexing framework (Channel bandwidth = 40MB/S)	151

List of Symbols

$arRCE$	Average relative rate control error
B_r	Remaining bits
B^T	Bit budget
c, c_1, c_2	Small constant to stabilize the division with weak denominator
$D(k, i)$	Content variation in the k th frame of the i th GOP
(d_x, d_y)	Motion vector
$f_{l,j,i}$	Performance of SVC at temporal level l , spatial level j and SNR level i
f_r	Frame rate
g_k	Golden-section point at iteration round k
$m(k, i)$	Number of Intra MBs in the k th frame of the i th GOP
$\bar{m}(k, i)$	Average of $m(j, i)$ ($j = 1, \dots, k$)
N	Counter
$n(i)$	Number of frames in i th GOP
Q	Perceptual quality
$\bar{Q}$	Averaged perceptual quality
QP^{qp}	Quantization parameter used to quantize residual
QP^{mv}	Quantization parameter used in motion estimation
r	Position on bitstream
R	Bit rate
R_A	Bit rate for auxiliary layer
R_c	Channel bandwidth
R^*	Optimal rate allocation
$R_{i,j}$	Bitrate allocated to the j th GOP in video i
R_{ref}	Bitrate allocated to reference view of the stereo video
R_{aux}	Bitrate allocated to auxiliary view of the stereo video
S_U	Upsampling operation
S_D	Downsampling operation
S_i	Weighted energy sum of the remaining MBs in frame i
SCX	Spatial complexity
TCX	Temporal complexity
t	Processing time
$T(k, i)$	Keyframe detection threshold for the k th frame of the i th GOP
T_{LO}	Lower bound
T_{UP}	Upper bound

W	Buffer fullness
X_1, X_2, K, C	Model parameters
$z_k(p)$	Luminance value at pixel position p in frame k
Ψ_{QP}	A quantization process followed by an inverse quantization process
λ	Lagrangian multiplier
ω	Weighting factor
$\rho(dx, dy, \delta_x, \delta_y)$	A set of candidate motion vectors around (dx, dy)
$\varphi(l)$	Most important resolution at temporal level l
μ_m	Average of m
σ_m^2	Variance of m
σ_{mn}	Covariance of m and n
ϵ	A predefined small value

Abbreviations

AVC	Advance Video Coding
CGS	Coarse Granularity Scalability
DCT	Discrete Cosine Transform
FGS	Fine Granularity Scalability
GOP	Group Of Pictures
HB	Hierarchical-B
HVS	Human Visual System
IBA	Implicit Bit Allocation
IMB	Intra MacroBlock
MAD	Mean Absolute Difference
MGS	Medium Granularity Scalability
MB	Macroblock
MC	Motion Compensation
ME	Motion Estimation
MSE	Mean Squared Error
MV	Motion Vector
MVF	Motion Vector Field
NAL	Network Abstraction Layer
PSNR	Peak Signal to Noise Ratio
QP	Quantization Parameter
RD	Rate Distortion
RDO	Rate Distortion Optimization
ROI	Region Of Interest
SAD	Sum of Absolute Difference
SATD	Sum of Absolute Transformed Difference
SNR	Signal to Noise Ratio
SSIM	Structure SIMilarity Index
SVC	Scalable Video Coding

Chapter 1

Introduction

1.1 Background and Motivation

Digital video becomes a part of our daily lives because of the advances in video coding technology and standardization [1] - [6] as well as the rapid development of network infrastructures and the exponential growth of computing power. The demands for multimedia services are rapidly increasing and the expectation on the quality of these applications is getting higher.

Because the storage capacity and bandwidth were and still limited resources, the main objective of video coding is compression. Over the last few decades, both the International Telecommunication Union-Telecommunication Standardization Sector (ITU-T) and the International Standards Organization (ISO) have been working on video compression standards that utilize waveform-based coding techniques to trade off the coding efficiency and the reconstructed video quality. With the introduction of H.264/AVC [6], significant improvements have been demonstrated on rate-distortion performance as compared with earlier codecs [1]

- [5]. The H.264/AVC could provide 50% bit-rate savings when compared with MPEG-2 [3], which is the most common standard used for digital television and DVD, and the compression gain of H.264/AVC over H.263 is in the range of 25% to 50% for different type of applications [7].

Besides the coding efficiency, scalability at a bitstream level is an attractive feature to transmit video over heterogeneous networks with varying client display resolutions, transmission bandwidths and network conditions. The Joint Video Team (JVT) of the ITU-T VCEG and the ISO/IEC MPEG has standardized a Scalable Video Coding (SVC) extension of the H.264/AVC standard [9]. The SVC design inherits many components of H.264/AVC and the base layer of SVC is basically coded in compliance with H.264/AVC. Any decoder that is conformed to H.264/AVC can decode the base layer of SVC. New tools have been developed in SVC to provide scalability on spatial and quality dimensions. To get a sub-bitstream representing a reduced spatial and/or temporal resolution, bitstream truncation is performed by discarding Network Abstraction Layer (NAL) units or network packets from an overall SVC bitstream [8].

Bit allocation plays an essential role in video coding. It is always a challenge problem to distribute the bitrate budget in a way such that the limited resources can be fully utilized while achieving high coding efficiency as well as optimized visual quality. Bit allocation is usually studied as an optimization problem. The optimization problem is formulated differently for various applications and the given bit-rate budget would be allocated among different units using different problem formulations. In general, all the applications can be categorized into 4 scenarios with different levels of bit allocation:

(1) **"One video source to One user"** - bit allocation of texture information

A single-layer codec, e.g. H.264/AVC, is usually used in this scenario, where the video

is encoded for the requirement of a single user. Bit allocation in this problem is more commonly referred to as rate control. The main purpose of rate control is to regulate the output bitrates through QP adjustment to achieve the best utilization of the available channel bandwidth and maximize the coding efficiency. Rate control involves bit allocation on GOP level, frame level and MB level. However, a fixed GOP size is usually adopted without considering the correlation between frames in majority of existing rate control algorithms. Temporal redundancies among frames cannot be properly removed with the fixed GOP size which also causes a degradation in coding efficiency. Thus it is necessary to study rate control scheme with adaptive GOP size and the problem on how to properly determine the GOP size based on video content is also worth considering.

- (2) **"One video source to Multiple users"** - bit allocation between texture information and header information

In heterogenous environment, the video source is encoded once by SVC at the highest required resolution and bitrate, while the resulting scalable bitstream can be extracted and decoded partially to serve many users with varying display resolutions, transmission bandwidths and network conditions. The layered structure of SVC makes the bit allocation problem to be a cross layer optimization problem. To optimize a specific layer according to user composition, both bit allocation between layers and bit allocation within each layer need to be studied. The latter also refers to implicit bit allocation which controls the tradeoff between motion information and residual information within each layer. If the bit allocation among current SVC layers is unbalanced in some situations, it is also worth studying how to optimize the coding structure in order to improve the coding efficiency.

- (3) **"Multiple video sources to One user"**- bit allocation among multiple videos

In broadcast applications, multiple videos are available for a user to watch at any time instance. Equally distributing the channel bandwidth and coding of each video independently cause large quality variations among videos as well as within each video. On the other way round, the limited channel bandwidth should be dynamically allocated among videos according to their relative complexity in order to achieve more uniform picture quality. It is therefore important to develop efficient algorithms to allocate the bandwidth among videos and to control each encoder accurately for a better utilization of the channel bandwidth.

(4) **"Multiple video sources to Multiple users"** - bit allocation among multiple videos with different spatial resolutions

In applications such as videoconferencing, multiple participants can be seen on screen at the same time. The participants involved in videoconferencing may have different display resolution, transmission bandwidth or network conditions. Videos captured by different participants can also be different in terms of resolutions, quality and etc. Therefore, each video should be encoded in a scalable format so that different portions of the scalable bitstream can be extracted accordingly by others. Meanwhile, the limited bandwidth needs to be distributed among multiple videos so that multiple extracted bitstreams can be decoded to maximize the overall quality. Many research problems on how to efficiently encode multiple videos as well as how to allocate bandwidth among multiple videos are involved in this kind of applications.

As illustrated in the application scenarios, bits are needed to be allocated to different units and thus the problems formulations are also different across different applications. Starting from distributing texture bits among GOPs in a single layer video, to the case of allocating bits between texture and header, and then to the case of channel bandwidth distribution among

several videos, efficient bit allocation techniques are indispensable to provide user with the best visual quality under the given network conditions, which is also the ultimate objective of video coding.

1.2 Objective and Main Contribution

This thesis develops algorithms to address bit allocation problems in the first three types of video applications mentioned in Section 1.1. These include bit allocation among videos, layers, GOPs, MBs as well as the bit allocation between motion information and residual information. On top of algorithm design, coding structure design is also studied to make the whole system works better with the proposed algorithms. The goal is to provide an efficient way to utilize the available resources so as to produce the best reconstructed visual quality with the given resource constraint. The fourth scenario mentioned in Section 1.1 will be studied in future work.

The contributions first cover texture bits allocation in a single video, which is also known as rate control. After taking care of texture bits allocation, research focuses on the bit allocation between texture bits and header bit in a single video for SVC. And in the last part, bit allocation among multiple videos is addressed. The main contributions of this thesis are summarized as follows:

- Designed a dynamic threshold model for keyframe detection. Every detected keyframe represents a start of a new GOP so that the sizes of GOPs are adaptive to video content. Designed a rate control scheme with adaptive GOP level bit allocation based on the proposed keyframe detection. The GOP size is an additional output of the rate control

scheme. An optimal GOP size can be obtained to be both content adaptive and at the same time achieves a good tradeoff between coding efficiency and random access.

- Proposed a new type of ROI in SVC for the combined spatial and SNR scalability. Implicit bit allocation, i.e., tradeoff between motion information and residual information, on each layer was studied. Designed a combined spatial and SNR scalability scheme, which is adaptive to customers' profiles by incorporating the concept of new ROI.
- Designed an improved inter-layer prediction scheme that uses two reference layers to achieve good balance among all layers for the combined CGS and spatial scalability. Proposed a new concept of auxiliary layer to improve the balance of bit allocation among layers. Both spatial redundancy and SNR redundancy between layers can be simultaneously removed and the overall coding efficiency is significantly enhanced.
- Designed a new and more accurate video complexity measure to determine base layer coding rate of multiple videos in broadcast applications. The proposed complexity measure is based on a new similarity index which incorporates characteristics of HVS. Designed a 2-D channel bit-rate allocation scheme to ensure a comparable quality level for all videos in broadcast applications while minimizing the quality fluctuations within each video.
- Designed a scalable stereoscopic video coding scheme to provide scalability and backward compatibility. Used the proposed scalable stereo video coding scheme to study joint bit-rate allocation for next-generation broadcast applications where multiple 3D stereo videos are transmitted simultaneously. Designed a bitrate allocation scheme that does adaptive inter-view bit allocation in each stereo video to achieve higher perceptual quality, and joint bit allocation among multiple stereo videos to obtain almost uniform perceptual qualities among all videos.

1.3 Outline of The Thesis

Chapter 2 gives an overview of the basic ideas behind the hybrid video coding structure. Development of International standards on video coding is then introduced. H.264/AVC and its scalable extension SVC, which are the two platforms used in the simulations of this thesis, are reviewed in detail. Rate control (i.e. bit allocation within a single video) and joint rate allocation are also briefly described.

A new adaptive scheme to perform GOP level bit allocation in rate control is proposed in Chapter 3. A dynamic threshold based scene change detection scheme is first proposed to determine the adaptive GOP size based on video contents. A new method to properly determine the optimal GOP sizes in rate control is then studied with the aim to get a good trade off between coding efficiency and random access.

In Chapter 4, three schemes that bring efficient bit allocation to SVC are presented. A combined SNR and spatial scalability scheme is first proposed to provide SVC encoder the flexibility to optimize a specific layer according to customer composition. The optimization is done by a proper bit allocation between motion information and residual information cross layers. A motion refinement scheme is then proposed to provide finer scalable granularity of that specific layer if the scalable bit rate range from the base layer to the specific layer is wide. An improved inter-layer prediction scheme is proposed to enhance the coding efficiency of layers at high resolution and a new concept of auxiliary layer that works with the new inter-layer prediction scheme is introduced to solve existing unbalanced bit allocation.

Chapter 5 studies the joint bit allocation for multiple videos in broadcast applications. A new statistical multiplexing system is presented, where SVC is applied to compress the multiple videos so that bit rate control of each video can be achieved by simple bitstream extraction.

The minimum guaranteed channel bandwidth is allocated among videos according to their relative complexity . A new complexity measure that incorporates the characteristics of Human Visual System (HVS) is introduced for this purpose. The proposed system not only minimizes the quality variations between videos, but also keeps the quality fluctuation small within each video. As 3D video is seen as a new type of media for next generation broadcast and streaming applications, bit allocation between multiple stereo videos is also investigated in this chapter. The problem is more complicated as it includes both bit allocation between videos as well as between views. The proposed framework does asymmetric bit allocation between views to obtain higher perceptual quality while the variance of perceptual quality among different stereo videos is also reduced as compared with existing method.

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

Chapter 2

Overview of Video Coding and Bit Allocation

2.1 A typical video compression system

A video is considered as a sequence of rectangular frames (or pictures) [10]. Statistical analysis of video signals indicates that there is a strong correlation both between successive frames and between pixels within the frame. Video compression could be realized by removing the redundant information from the video signal. If lossless compression schemes are employed, the statistical redundancy will be removed so that the original signal can be reconstructed perfectly at the receiver end. Unfortunately, the compression ratio achieved using lossless methods cannot satisfy the application requirement. As the Human Visual System (HVS) is not sensitive to loss of certain spatio-temporal visual information, lossy compression techniques can be used to adjust the tradeoff between video bit-rate and image quality.

FIGURE 2.1: Generalized block diagram of a video compression system.

There are several types of redundancies existed in the video signals. These redundancies are either spatial due to the similarities of neighboring pixel values within the same frame, statistical because of the possibility of occurrence of intensity levels within the video sequence or temporal as a result of the similarities encountered amongst adjacent video frames. Figure 2.1 plots a generalized video compression system. A number of techniques will be applied on each input video frame during the compression process. The typical compression techniques are the four listed below:

1. Prediction:

Prediction is a process to estimate the values of the input samples such that a set of possible values is formed. Normally the encoder will indicate how the prediction is created and the types of prediction can be selected based on analysis of the input samples in the system. Prediction typically operates on macroblocks, which are square-shaped blocks of neighboring pixels. These blocks in one frame are compared with some or all of the possible blocks in a predefined search range of another frame. Only differences between these blocks and their best-matches in another frame are coded and transmitted.

2. Transform:

Transformation is a process that maps the pixels into a transform domain prior to data

reduction. A transformation can capture the essence of the input signal by using frequency analysis and it prevents repeatedly representing similar values. The strength of transform coding in achieving video compression is that the image energy of most natural scenes is mainly concentrated in the low frequency region, and hence into a few transform coefficients. These coefficients can then be quantized with the aim of discarding insignificant coefficients without significantly affecting the reconstructed image quality. The Discrete Cosine Transform (DCT) [20] is broadly used because it has smoothly varying basis vectors that resemble the intensity variations of most natural images, such that image energy is matched to a few coefficients.

3. Quantization:

After transformation, a significant part of the image energy is concentrated at the lower frequency components, while less for the high frequency ones. The human eye is less sensitive to picture distortions at higher frequencies. During quantization, each component in the frequency domain is divided by a constant and then rounded to the nearest integer value. As such, high frequency information can be rounded to zero or very small values. Hence a lot of high frequency components can be discarded to save the coding bits.

4. Entropy coding:

Entropy coding is to convert a set of symbols into a compressed bitstream suitable for transmission or storage. Input symbols may include quantized transform coefficients, motion vectors, markers, headers and some supplementary information. Variable Length Code (VLC) is a well-known entropy code. In VLC, frequently-occurring symbols are represented with short code words whilst less common symbols are represented with long code words to achieve compression of the data.

FIGURE 2.2: Development of various video coding standards.

2.2 Video Coding Standards

Despite increasing communication capacity and storage space, the demand for better compression techniques has not diminished as higher video resolution requirement grows. Over the last few decades, both ITU-T and ISO have been working on video coding standards that serve as languages through which a video encoder and decoder communicate. Figure 2.2 illustrates the development of various video coding standards. A video standardization process defines the syntaxes and semantics of a compliant video bitstream, ensuring that all compliant decoders can understand an encoded file produced by any compliant encoder.

2.2.1 Earlier Codecs

Although numerous tools have been developed that led to improvement of modern codecs over earlier ones, the core technology behind the codecs remains similar. Motion compensated prediction to remove inter-frame redundancy followed by transform coding for energy compaction that allows effective quantization has been proven to be an exceptionally effective method to compress video data.

The success of hybrid video encoders can be attributed to several key components:

- Motion compensated predictive coding which reduces redundancies that exist between temporally distinct video frames. Predictive coding reduces the power of the symbols that are coded; fewer bits are required to code prediction residues compared to actual pixel values.
- Transform coding of prediction residues leads to coefficients that are less correlated than the original samples; adjacent samples typically have substantial correlation and separate quantization of each would be an inefficient way to encode them. In addition, transformation leads to energy compaction such that information is concentrated in only a few transform coefficients. As a result, they can be quantized and encoded more efficiently.
- Entropy coding of quantized transform coefficient reduces the number of bits required to represent them. The goal of entropy coding is to reduce the average number of symbols sent with no loss of fidelity in the decoded signal.

2.2.2 H.264/AVC

For a number of years, work on the "Advanced Video Coding" standard, now known as ITU-T Recommendation H.264 and as ISO/IEC 14496 (MPEG-4) Part 10 [6] occupied the video coding community. Even after the standardization work is done and coding standard specified, the H.264/AVC video coding scheme continues to be an active area of research. Due to the excellent coding performance of the codec, both the research community and industry players expect it to be widely adopted.

H.264/AVC has many new features that improve picture quality and compression performance [7]:

- In-loop deblocking that removes artifacts caused by block-based prediction and transform improves both perceptual quality of reconstructed pictures and coding performance [11].
- H.264 allows the use of smaller block sizes during motion compensation. The use of finer partition during the motion compensated prediction improves the prediction accuracy and reduces bits required to code the residue. This can potentially lead to better coding performance [12].
- Weighted prediction allows prediction signal to be weighted. This significantly improves the coding of video frames containing fades [13].
- Improved SKIP mode allows motion vectors to be derived from neighbouring block despite expending no bits for coding motion vectors [14].
- Spatial prediction for intra frames reduces the variance of data that has to be coded for intra coding. Edge directed prediction removes redundancy within a picture effectively, resulting in fewer bits needed for coding the picture.
- The use of arithmetic coding, adaptive codes and context modelling [15].
- A DCT-like integer transform that resulted in an exact definition of the inverse transform [16].
- Multiple reference frames can be used for motion-compensated prediction.

2.2.3 Scalable Video Coding Extension of H.264/AVC

Scalable video coding has been an active research area for decades. The earlier video coding standards, such as MPEG-2 Video/H.262 [3], H.263 [4] and MPEG-4 Visual [5] already include

several tools to support scalability on temporal, spatial and SNR dimensions. However, none of the scalable profiles of these standards has been really used. This is mainly because scalability functionality always came along with a significant loss in coding efficiency and a large increase in decoder complexity as compared with the corresponding non-scalable profiles. Moreover, there are alternative methods to provide scalability functionality. Simulcast is the most straightforward method to transmit two or more non-scalable stream without removing any redundancy in the bitstream, as shown in Fig 2.3 (a). Simulcast is simple to implement but suffers from a significant increase in bitrate. A single bitstream can also be adapted to serve various clients in heterogeneous network through transcoding, which is currently used for streaming service in 3G systems. Hence, a scalable video codec has to compete against these alternatives.

The ITU-T VCEG and the ISO/IEC MPEG has jointly standardized a Scalable Video Coding (SVC) extension of the H.264/AVC standard. Through SVC, an encoded video file can be decoded to fulfill several frame-rate, resolution or quality requirements after appropriate extraction. The basic SVC design can be classified as layered video codec. Enhancement layers successively refine their base layers, resulting in decoded videos that are of higher resolution, higher frame rate or better quality. An SVC encoder with three spatial layers is illustrated in Figure 2.4. The main challenge of scalable video coding is to remove inter-layer redundancy. Ideally, the coding performance of scalable video should approach that of a single layer codec when producing decoded pictures of the same quality. This is usually not the case. Although the SVC outperforms the scalable profiles of earlier standards and simulcast, the coding efficiency is still shown to be lower than the single layer coding [17]. Therefore, the flexibility of being able to represent videos of different resolution, frame rate and quality comes at a cost of a coding performance penalty.

FIGURE 2.3: (a) Separate encoding/transcoding and simulcast transmission. (b) Scalable video coding and transmission

There are two different ways to enable scalability in a codec. One is to use layered approach, which has appeared in many previous standards such as [3], [5]. The other one is to use a technique that is intrinsically scalable, such as bitplane arithmetic coding. SVC adopts a combination of the two approaches to enable a full spatio-temporal and quality scalable codec.

FIGURE 2.4: Overall SVC encoder structure with three spatial layers.

2.2.3.1 Temporal Scalability

Temporal scalability is defined to represent the same video content in different temporal resolutions or frame rates. When temporal scalability is a requirement in coded sequence, the video encoder has to avoid using frames of higher temporal levels for prediction when encoding a particular frame. This can be achieved by encoding each Group of Pictures as a hierarchy of frames. As shown in Figure 2.5, the temporal scalability is enabled by the concept of hierarchical B pictures. The video frames in the lowest temporal level form the base layer and each subsequent temporal level becomes an enhancement layer that enhances the frame rate of the video. B pictures (Bi-directional pictures) are pictures that are coded using pictures in both temporal directions as reference frames. Blocks in a B picture have the choice of using either past frames or future frames (or a combination of both) during motion compensated prediction. A newly uncovered object in the current coded frame cannot be successfully predicted with past frames; this will inevitably affect prediction and coding performance. This problem can be partially alleviated with forward prediction or bi-prediction. Hence, the inclusion of B pictures usually brings coding performance improvement. When encoding for

FIGURE 2.5: Hierarchical-B coding structure (GOP size = 8, No. of temporal levels = 4) [8].

low-delay applications, temporal scalability can also be achieved with hierarchically-arranged P-frames.

2.2.3.2 Spatial Scalability

Spatial scalability is defined to represent the same video content in different spatial resolutions or frame sizes. Similar with many earlier standards, SVC also utilizes the multi-layer coding mechanism to realize spatial scalability. Different spatial layers have different spatial resolutions. The lowest resolution video data is referred to as the base layer, and the higher resolution video data is referred to as the enhancement layer. In each spatial layer, both the hierarchical motion compensated prediction (MCP) and the intra prediction are implemented as for single-layer H.264/AVC coding.

To further improve the coding efficiency as compared to sending videos with different spatial resolutions by simulcast, inter-layer prediction mechanism is employed to explore the correlation among different spatial layers as shown in Figure 2.4. The inter-layer prediction aims to removes redundancy among layers and it is realized by reusing some of the encoded information from lower resolution video sequence to code the corresponding higher resolution video sequence. The prediction has been made adaptive, so that an enhancement layer has the

freedom to choose any reference layer from lower resolution layers. There are three inter-layer prediction techniques introduced in SVC [8]:

Inter-layer motion prediction

As compared to H.264/AVC, two more macroblock modes, a "*base layer mode*" and a "*quarter pel refinement mode*" [8], are introduced to make use of the motion information in the lower spatial layer for prediction of the motion vectors in higher layers. For both of these two modes, the macroblock partitioning is obtained by up-sampling the corresponding co-located 8×8 block in the reference layer. The reference index for the up-sampled macroblock partitions are the same with that for the co-located blocks in reference layer. The corresponding motion vectors are derived through scaling the motion vector in the reference layer by a factor of 2. For macroblock with base layer mode, no additional motion information is coded or transmitted, while for macroblock with quarter pel refinement mode, a quarter-sample motion vector refinement is further transmitted for each motion vector. Meanwhile, the up-sampled motion vector from the reference layer could also be used as motion vector predictor for the macroblock in the enhancement layer. If none of the above techniques are used, the macroblock modes as well as the corresponding reference indices and motion vector differences are encoded according to the H.264/AVC syntax.

Inter-layer residual prediction

Inter-layer residual prediction can be used for coding of each inter-coded macroblock. If the motion vector of a block in the current layer is the same or similar to that of the corresponding block in the lower layer, it is preferable to use the residual coded in the lower layer as reference for prediction of the current layer residual. By coding only the difference between the residual in two layers, the compression efficiency can be significantly improved. However, when the motion vectors are not similar it is unlikely that a prediction of the residual signal could

improve the coding efficiency. Therefore, the inter-layer residual prediction is adaptive and a flag is sent for each inter-coded macroblock.

Inter-layer intra prediction

The inter-layer intra-prediction can be applied to a macroblock whose corresponding block in the base layer is intra-coded. With this mode, the reconstructed texture information from the lower spatial layer is up-sampled using the 6-tap filter defined in H.264/AVC to predict the texture information in higher layer. The prediction residual is coded and transmitted as in H.264/AVC. In this mode, the intra prediction signal can be simply attained by de-blocking and interpolating the corresponding 8×8 luminance block in the lower layer. Therefore, the decoding complexity is greatly decreased.

2.2.3.3 Quality Scalability

Quality or SNR scalability is defined to represent the same video content in different perceptual quality. With quality or SNR scalable coding, the input video can be compressed into multiple layers including a quality base layer and one or more quality enhancement layers. Figure 2.6 illustrates a general structure of a scalable bitstream with a base layer and L quality enhancement layers. For the quality base layer, H.264/AVC based transform coding is carried out. The quality base layer provides the minimal visual quality with an initial quantization step size. To achieve the quality enhancement layers on top of the quality base layer, different scalable coding approaches are employed to provide different granularity of scalability as follows.

Coarse-Grain Quality Scalability (CGS)

Coarse-Grain Quality Scalability (CGS) can be seen as a special case of spatial scalability without change on frame sizes. In CGS, the inter-layer prediction employs the same mechanisms

FIGURE 2.6: General structuring of a scalable bitstream with a base layer and L quality enhancement layers.

used in spatial scalable coding. However, there is no need to have up-sampling and inter-layer de-blocking for intra-coded macroblocks in reference layer. When realizing the inter-layer prediction for CGS, texture information is usually refined in the enhancement layer by re-quantization of the residual texture signal with a smaller quantization parameter relative to that used for the previous CGS layer.

A number of quality or SNR enhancement layers can be generated for CGS. However, the CGS only supports limited truncation points in a scalable bitstream, where the number of truncation points is equal to the number of layers. Another limitation of CGS is that the enhancement layer has to be completely transmitted or decoded to refine the picture quality. If the available bandwidth is not sufficient for transmission of the enhancement layer, the server will drop the entire enhancement layer for rate adaptation, which may lead to under-utilization of the channel bandwidth.

Medium-Grain Quality Scalability (MGS)

While CGS provides quality scalability by dropping complete enhancement layers, SVC provides a finer granularity level of quality scalability by partitioning a given enhancement layer into several quality layers. This quality scalable mode is referred to as medium-grain quality

scalable (MGS) coding. Those quality layers inside the enhancement layer are identified by a "*quality identifier*", which increases with every additional quality refinement representation. In MGS, a given enhancement layer of a given video frame can be split into up to 16 MGS layers. The prediction modes and the motion vectors may be modified between successive MGS layers.

The structure of motion compensated prediction (MCP) for MGS is designed differently in comparison to earlier scalable codec, such as fidelity scalable coding in MPEG-2 Video [3] and fine-grain scalability coding in MPEG-4 Visual [5]. The MCP for MPEG-2 Video is done by only using the reconstructed reference with highest available quality. Any loss of a quality refinement packet will cause drift between encoder and decoder and the propagation of this drift can only be stopped at Intra-coded frame. For fine-grain scalability coding of MPEG-4 Visual, in order to completely avoid drift, only the reconstructed base layer pictures are used for further prediction that gets low coding efficiency as a result. In MGS design, the reconstructed reference with highest available quality is used for MCP of non-key pictures, while only the key picture is predicted from the reconstructed base layer reference. Therefore, the drift between encoder and decoder cannot be propagated beyond the key pictures and at the same time, high coding efficiency can be achieved in non-key pictures without increasing the complexity.

Fine-Grain Quality Scalability (FGS)

In order to support finer granularity, so-called progressive refinement (PR) slices have also been introduced. Within CGS enhancement layer or spatial enhancement layer, fine-grain quality scalability (FGS) is achieved by refining the transform coefficient successively, starting with a base quality layer obtained by the H.264/AVC compatible base layer coding. This is done by repeatedly decreasing the quantization step size (QP increase of 6) and applying

a modified entropy coding process akin to sub-bit-plane coding. It is possible to flexibly truncate the progressive refinement of FGS at any point, so that the decoded video can be improved in a fine-granular way. It should be noted that FGS coding mode has been removed from the SVC amendment that has been finalized in July 2007.

2.3 Rate Control for Video Coding

Rate control plays an important role in compressed video communications. It regulates the output rate of the coded bitstream in order to achieve the best video quality under the channel bandwidth constraint. It can be considered as a bit allocation problem in a single video. Although video coding standards have defined the syntax of video decoding, the issue of designing the rate controller for an encoder is left open for optimizations.

As described before, video compression is achieved by reducing both the temporal redundancies and spatial redundancies within the source video data. Since these two types of redundancies are time-varying in real-world video sequences, if there is no rate control mechanism employed at the source coder side, i.e., a fixed QP is used during the encoding process, the output bit rate of the compressed bitstream will also vary in time. For example, the output bit rate will increase at the scenes which contain complex motions (with higher temporal activity) while fewer bits will be generated when coding more stationary scenes (with lower temporal activity). This relationship between temporal activity and output bit rate is illustrated in Figure 2.7. Figure 2.7(a) illustrates the level of temporal activity of a sequence along time. As can be seen in Figure 2.7(b), if a constant QP is used throughout the coding process, large fluctuations in the output bit rate will be observed while constant visual quality

can be achieved for the compressed video. As the output bit rate to the channel is varying, this kind of coding scheme is referred to as variable bit rate (VBR) encoding. The VBR video encoders are suitable for the applications where very large buffers and fast channels are available. However, for realtime low-delay video communications, the constant bit rate (CBR) encoding is more desirable as depicted in Figure 2.7(c). In CBR scheme, the buffers have to be put between the encoder and the CBR channel to smooth out the variation of the output compressed bitstream. Given the buffer and channel conditions, the rate controller should adaptively change the QP for the current coding unit (a frame or a MB) such that the bit rate of the compressed video data can comply with the channel bandwidth. Moreover, the buffer fullness should be carefully controlled to avoid overflow and underflow. Because the undesirable frame skipping as a result of buffer overflow will cause motion jerkiness and on the other hand, the channel bandwidth will be wasted if buffer underflow happens. Besides regulating the output bit rate according to the buffer and channel status, another objective of rate control is to achieve stable and better reconstructed video quality by adaptively allocating bit budget for different coding units.

Normally, the rate control process can be segmented into the following stages as shown in Figure 2.8. At the first step, the target bit allocation for the current coding unit is performed. In this stage the coding unit can be a GOP in the video sequence, a frame in the current GOP or a MB in the current frame. The bit allocation can be done not only based on the channel bandwidth and the current buffer status but also by taking the temporal complexity into consideration [30], [31]. As a result, more consistent video quality can be achieved with less quality fluctuation. In addition, region based bit allocation methods [27] - [29] can allocate more bits to the region-of-interest (ROI) within a video frame, such as the foreground, while fewer bits are used to code the background in order to achieve better subjective visual quality.

FIGURE 2.7: VBR and CBR coding schemes with corresponding relations between the output bitrate and quality.

After bit allocation for the current coding unit, the rate controller should choose an appropriate QP for the quantizer using a certain rate-distortion (R-D) model. Ideally, if we know the exact R-D characteristics of each coding unit we can select the optimal QP thus the output bit rate of this coding unit will match its target bit budget perfectly. However, the actual R-D characteristics are not available without any pre-encoding process, mathematical R-D models are widely used to estimate the QP based on previous coding statistics [21] - [25]. The current coding unit will be coded using the QP determined by the R-D model. At last, the R-D model and buffer status are updated according to the actual coding statistics.

FIGURE 2.8: Procedures for a typical video rate control scheme.

2.4 Joint Rate Allocation for Statistical Multiplexing

The advances in digital video compression and transmission technology have also enabled delivery of several digitally compressed video programs in a channel, which was used to transmit a single analog program. Multiple video programs are usually encoded individually, then multiplexed and transmitted over a bandwidth-limited network to receivers for video decoding and display.

The simplest approach to distribute channel bandwidth among different videos is equal rate allocation, which means each video is coded at an equal and constant bit-rate. The rate control must be carried out on each video to regulate its output bit-rate as shown in Figure 2.9. However, in a multiplexing system, different videos may have different scene activities, and within each video, the scene activity may change dramatically over time. According to the rate-distortion theory [86], equal rate allocation may lead to uneven distortions between different videos as well as within each video due to variation of scene complexity. This system has another drawback, which is the channel bandwidth cannot be fully utilized due to the limited accuracy of the rate control scheme.

To achieve uniform video quality for various videos, the channel's available bandwidth should be dynamically allocated to different videos according to the relative complexity of each video source. This is called statistical multiplexing [69]. Figure 2.10 illustrates the block diagram

FIGURE 2.9: Independent coding of multiple video signals.

FIGURE 2.10: Joint coding of multiple video signals.

of a statistical multiplexing system with joint rate allocation. An accurate estimation of each video's complexity model is a prerequisite for efficient rate allocation. In addition, video encoder also needs to be accurately controlled in order to use the channel bandwidth more efficiently. In general, the joint rate allocation in statistical multiplexing should ensure that [32]:

- all videos should be encoded to satisfy a minimum quality constraint,
- all videos should be encoded with similar quality (constraint on "*inter-program fairness*") [69],
- for each video, the quality variation of the reconstructed video should be smooth with time (constraint on "*intra-program smoothness*") [33] [34].

Chapter 3

Adaptive GOP Rate Control Based On Keyframe Detection

3.1 Introduction

Bit allocation within a single video source is more commonly referred to as rate control. Even though rate control is a non-normative part of video coding standards, it plays a very important role in achieving best video quality under the channel bandwidth constraint. Therefore, several benchmark rate control schemes have been developed for various video coding standards, such as MPEG-2 TM5 [35], H.263 TMN8 [36], MPEG-4 VM18 [37][38], H.264/AVC adaptive rate control [39] and so on. Generally, there are two main goals of rate control. One is to guarantee that the actual bit rate of a video stream matches a given bit rate which can be constant [35] or time varying [39]. The other is to maximize the coding efficiency. The inputs of a rate control algorithm are a video sequence and a given bit rate. The outputs of existing rate control algorithms are the values of quantization parameters (QPs) and the number of

skipped frames [35]-[39]. It is noted that existing rate control schemes adopt fixed group of picture (GOP) sizes which periodically insert INTRA frames, also known as keyframes, as random access points without considering the contents in the video sequence. Having random access points in a compressed video bitstream is very necessary. In multi-channel service, it allows fast channel switching and in many other applications, it enables important functionalities such as pause, fast forward, reverse, and fast reverse access. However, a possible disadvantage of such a fixed GOP structure is that the temporal redundancy among frames cannot be well removed and thus the GOP level bit allocation is not effective for sequences with varying scene complexity from one GOP to another.

If a frame in a GOP contains abrupt scene changes, fast movements and etc., which results a large part of the frame cannot be found in the reference frames, the prediction efficiency of the GOP would be very low and it would degrade both the subjective and objective quality of the video in order to match the given bit rate. On the other hand, if a frame with strong correlation to previous reference frames is selected as a new keyframe (i.e., start of a new GOP), the coding efficiency would also drop because extra bits will be consumed on the frame due to temporal redundancy. For the GOP consisting of the group of strongly correlated frames, all the frames can be well predicted from reference frames and visual quality will be greatly increased instead. As a result, the video quality varies dramatically from one GOP to another due to the varying contents in each GOP. It is thus reasonable to believe that the best keyframes to encode a video are the frames that have low temporal correlation inside the video. Properly selecting the keyframes and making the scheme fast and robust is always an attractive topic in video coding optimization. While at the same time, for rate control, if we incorporate a good keyframe detection scheme to make the sizes of GOP, which is decided by the locations of keyframes, to be content adaptive, great improvement on coding efficiency

can be expected too. In [40], an adaptive GOP scheme was applied to INTRA frame rate control where the GOP size was adjusted according to the content of video and a predefined maximum GOP size. There might be problems if the maximum GOP size is not properly set. If the maximum GOP size is too large, the lack of sufficient random access points usually cause problems in synchronization, error propagation etc., whereas a too small maximum GOP size results in frequent INTRA frame insertion that consumes unnecessary bandwidth. Therefore, it is desired to have a scheme to properly determine the sizes of GOPs in rate control that is both content adaptive and at the same time achieves a good tradeoff between coding efficiency and random access.

In this chapter, a dynamic threshold model for real-time keyframe selection is first proposed. The proposed algorithm is based on a video content variation measurement that takes advantage of the information that is already generated during the encoding process of each frame. In particular, it is relied on the number of intra macroblocks (IMBs) used to encode an INTER frame as in [50] based on the following two observations: 1) the number of IMBs is large for a high motion video sequence; 2) the number of IMBs is large when an abrupt change occurs. The proposed dynamic threshold, which captures the characteristics of both scene changes and fast motion, is updated adaptively with the variation of IMBs, the resolutions and the quantization parameters (QPs) of the video sequence. The way of calculating the proposed threshold is very straightforward and flexible, whereas the method in [50] is constrained by predefined parameters that are determined under some specific encoder configurations. Based on the proposed dynamic threshold model, a new method to determine the optimal GOP size in rate control is also proposed. Since an INTRA frame is a necessary random access point, the average value of the sizes of GOPs is taken into consideration and it is required not to be greater than the size of a fixed GOP in the proposed method. An upper

bound is also added to the sizes of GOPs for the function of random access. To avoid situations where there are too many INTRA frames, a lower bound is also added to the sizes of GOPs. The proposed method is able to overcome problems mentioned above on the scheme in [40]. And the experimental results show that the proposed method indeed achieves PSNR improvement.

The rest of this chapter is organized as follows. An overview of rate control schemes with fixed GOP size and some related work on scene change detection are presented in Section 3.2. In Section 3.3, the details of content variation metric, the dynamic threshold model and the keyframe detection algorithm are described. Section 3.4 presents the proposed rate control scheme with adaptive GOP size. Section 3.5 contains performance evaluation of the proposed scheme and Section 3.6 summarizes this chapter.

3.2 Background and Related Works

3.2.1 Rate Control with Fixed GOP size

Various rate control techniques have been proposed for video coding in the past decade. In this section, three well-known rate control algorithms that have been adopted by the international standards are briefly reviewed, including TMN8 [36] for H.263, VM18 [38] for MPEG-4 and adaptive rate control algorithm [39] for H.264/AVC. All the three algorithms adopt a coding structure with fixed GOP size. The target bits allocated to each GOP are determined by frame rate, the number of frames in each GOP and the available channel bandwidth. As discussed earlier in Section 3.1, the bit allocation is not effective if the GOP size is fixed without considering the characteristics of the video content.

TMN8 Rate Control Algorithm

In general, the TMN8 [36] rate control is performed at two layers: frame layer target bit allocation and MB layer QP selection. In the first step, the number of bits allocated to the current frame is determined based on the current buffer fullness, the frame rate and the channel bandwidth. The current buffer fullness can be calculated as

$$W = \max(W_{prev} + D - R/F, 0) \quad (3.1)$$

where D is the actual amount of bits used to encode the previous frame, W_{prev} is the previous buffer fullness, R is the channel bandwidth and F denotes the frame rate. Similar to TMN7 [52] rate control scheme, the TMN8 method skips encoding next frames when the buffer fullness W is larger than a predefined threshold T , ($T = R/F$ by default for low-delay applications). Note that, this frame skipping method is to guarantee that the buffer delay will not be larger than a frame interval. The frame bit budget for the current frame is further estimated as

$$B^T = R/F - \Delta, \quad (3.2)$$

where

$$\Delta = \begin{cases} W/F & \text{if } W > Z \\ W - Z \cdot T & \text{otherwise} \end{cases}. \quad (3.3)$$

in which Z is set to 0.1 by default. The MB layer rate control is to select QP for each MB within the current frame after the frame level bit allocation is done. Based on the RD model proposed in [53][54], the optimal quantization step size Q_i^* for the i th MB in current frame is

derived as following equation

$$Q_i^* = \sqrt{\frac{256K_i}{L_i} \cdot \frac{\sigma_i}{\alpha_i} \cdot S_i} \quad (3.4)$$

where σ_i is the standard deviation of the luminance and chrominance values in the MB, $S_i = \sum_{k=1}^N \alpha_k \sigma_k$ is the weighted energy sum of the remaining MBs in current frame. $L_i = \beta_i - 256N_iC_i$ is the number of remaining bits for coding the rest MBs in current frame excluding header bits. Note that, β_i and N_i represents the bits budget and the number of MBs that remain in current frame respectively. The model parameters K_i and C_i are updated adaptively after coding each MB [36]. The TMN8 rate control method performs better than the previous approaches for H.263 video in terms of the reconstructed video quality and the bit rate control accuracy.

VM18 Rate Control Algorithm

The VM18 [38] rate control algorithm designed for MPEG-4 video coding is based on the quadratic RD model [37] given by

$$R_n = S_n \left(\frac{X_{1,n}}{Q_n} + \frac{X_{2,n}}{Q_n^2} \right) \quad (3.5)$$

where Q_n is the quantization step size of current frame n , S_n denotes the coding complexity which is often substituted by the sum of mean absolute difference (MAD) of the residue frame. Based on the assumption that neighboring video frames of the same type may have similar R-D characteristics, the model parameters $X_{1,n}$ and $X_{2,n}$ are updated using linear regression method from previous parameters.

The target bitrate for the current frame is computed based on the bits available and the bits of the last encoded frame. Basically, more coding bits will be allocated to the current frame if the

bit rate of the previous frame is high. On the other hand, fewer bits will be allocated to the current frame if there are fewer bits left for encoding. The initial target bit rate is calculated by

$$B_n^T = \max \left(\frac{R}{F}, 0.95 \frac{B_r}{N_r} + 0.05 A_{n-1} \right) \quad (3.6)$$

where R and F are the bit rate and frame rate of the source video, respectively. B_r is the number of remaining bits for encoding this video, N_r is the number of remaining frames to be coded and A_{n-1} is the actual bit used for coding the previous frame. The initial frame bit budget from (3.6) will be further adjusted to avoid the buffer overflow and underflow [38]. When frame layer bit allocation is completed, the QP for the current frame can be calculated using equation (3.5) and it will be clipped between 1 and 31. Once the current frame is encoded, if the buffer status is above 80%, the next frame is skipped in order to prevent the buffer from overflowing.

Adaptive Rate Control for H.264/AVC

The concept of basic unit was introduced in the rate control scheme of H.264/AVC [39]. In this framework, the rate control can be implemented at different levels according to the definition of the basic unit such as frame level, slice level, MB level or a group of contiguous MBs. Similar to previous rate control algorithms, the target bit allocation is conducted based on both the channel bandwidth and the buffer status. After knowing target bit allocated, QP should be determined through R-Q model which requires the actual statistics such as the MAD or the residue variance. However, the actual statistics is only available after performing Lagrangian-function-based rate distortion optimization (RDO) in motion compensation process. QP must be known in order to calculate the Lagrangian multiplier λ in RDO. Therefore, this results in the so-called "chicken and egg dilemma" [39] in the rate control of H.264/AVC. To solve the

dilemma, a linear model for MAD prediction is used to work with the classic quadratic R-Q model [24] in the adaptive rate control scheme. Using frame level rate control as an example, firstly, the target bit budget for the current P-frame is determined by the average available bit budget and further modified according to the buffer status. Then the MAD of the previous P-frame is used to predict the MAD for current frame using a linear model. The model parameters are updated during the coding process. Finally, substituting this predicted MAD for the actual MAD in the quadratic R-Q model, the QP of current frame is thus obtained.

3.2.2 Scene Change Detection Algorithm

Many types of scene change detection schemes and adaptive GOP size detection schemes have been proposed in literature. All the schemes define a similarity measure between two consecutive frames. When this measure shows a big change that exceeds a threshold, a scene change or a keyframe is identified. Schemes in [41]-[43] used the evolution of motion vectors as the measure. This approach is able to find frames with complex motion but not always true for frames with abrupt scene change as scene change may not always result in large motion vector values. In addition, this approach is usually computationally intensive. Authors in [43]-[45] made use of the information such as sum of absolute difference (SAD) or sum of absolute transformed difference (SATD) in the similarity measure. They are very effective in detecting scene changes but frames with big motion may be easily missed. Li *et al.* [46] proposed a dynamic threshold for scene change detection based on bit rate fluctuations. The method can detect frames with either scene change or big motion and it is simple and fast. However, it can only be used for variable bit rate video system without enabling rate control. Usach *et al.* [50] proposed to make use the number of INTRA macroblocks as similarity measure. Although the threshold is adaptive to resolution and quality of encoded video, the calculation

of the threshold is limited to certain encoder configurations and thus lacks of generality. There are many other different methods [47]-[49], but they all have the disadvantage of high computational complexity which makes them not suitable for real-time applications.

3.3 Content Adaptive GOP Size Control

In this section, a dynamic keyframe detection scheme is proposed based on a newly defined variation metric and a new dynamic threshold function. The general principle is that, for every inter-coded frame, when the content variation exceeds the threshold, the frame is detected as a keyframe and thus re-encoded as an INTRA frame to improve the coding efficiency.

3.3.1 Video Content Variation Metric

When the correlation between a frame and its reference frame is low, differences between two frames due to fast motion or scene change cannot be efficiently predicted and compensated by inter modes, so intra prediction has to be used which leads to an increase in the number of IMBs in the P-frame. Thus the lesser the correlation between a frame and its references, the more IMBs are necessary to encode it. To show this behavior, Figure 3.1 shows the percentage of IMB used in each frame of an inter-coded video sequence concatenated by five different standard sequences from "Foreman", "Bus", "City", "Crew" and "Soccer" with 20 frames each at CIF (352x288) resolution. The sequence is encoded at 15fps with IPP...P structure on H.264 encoder [55]. It can be observed that almost 100% of the macroblocks are intra-coded on the frame where a new sub-sequence starts, and for frames with complex motion (such as in sequence "Crew"), more IMBs are used than low motion frames (such as in sequence "City").

FIGURE 3.1: Number of IMBs in each frame of a concatenated sequence encoded at CIF resolution and 15fps.

The number of IMBs in each frame is an effective measure of the content. A big advantage of using the number of IMBs is its simplicity. It can be obtained easily during encoding process without any extra complicated computations. In this work the content variation of frame k is defined as:

$$D(k, i) = m(k, i) - \bar{m}(k-1, i), \quad (3.7)$$

where $m(k, i)$ is the number of IMBs in the k th frame of the i th GOP and $\bar{m}(k-1, i)$ is the average value of $m(j, i)$ ($j = 1, 2, \dots, k-1$) in the i th GOP. The content variation is measured and compared with a predefined threshold for each frame. If the variation exceeds the threshold, a keyframe is detected.

3.3.2 Dynamic Threshold Model

If we want to preserve the real-time character of the keyframe selection scheme, we should only use the previous frame values without looking in the future when deciding the threshold.

For the proposed dynamic threshold function, local statistical properties of the sequence are used. Let $\sigma(k-1, i)$ be the standard deviation of the number of IMBs in the i th GOP. The threshold function for frame k in the i th GOP is a function of $\sigma(k-1, i)$ as shown below:

$$T(k, i) = \gamma \cdot \sigma(k-1, i). \quad (3.8)$$

Conventional approaches determine the threshold according to an empirical record without considering other parameters. While analyzing data from the training sequences, it is found that the number of IMBs as well as the mean and standard deviation on the IMBs strongly correlate with changes in resolution and QP. Therefore, we speculate that the resolution and QP values should constitute the main components within the threshold calculation. In the proposed dynamic threshold function, γ is thus not a constant but a function of both resolution and QP value of the sequence. In the following, the impact of resolution and QP value on the threshold is analyzed.

Correlation Between the Resolution and the Threshold

Several tests have been done on video sequences to evaluate the correlation between resolution and γ when QP is kept unchanged. The number of IMBs as well as the mean m_k and the standard deviation σ_k of the IMBs have been recorded for analysis. Similar behaviors have been observed from different testing sequences. In this section, the data obtained from one of the testing sequences is used for illustration. The sequence consists of 100 frames of "Crew" and 100 frames of "Soccer". It is coded as IPP...P structure by H.264 encoder [55]. Figure 3.2 shows the trend of changes in the mean and deviation of IMBs when resolution of the sequence changes from QCIF (176x144) to CIF (352x288) and then to 4CIF (704x576). Due to different number of macroblocks available in each frame at different resolution, the data are

plotted in percentage. It can be seen from the figure that, both m_k and σ_k of IMBs increase with a larger resolution. QP is set to 20 in the test shown in Figure 3.2 but similar behaviors are observed at other QP values.

FIGURE 3.2: (Left) Changes on mean of IMBs (Right) Changes on standard deviation of IMBs at different resolutions

Recall the behavior of IMBs discussed in Figure 3.1, keyframe candidates usually contain a large number of IMBs and the numbers are similarly large even in different candidates. So the number of IMBs $m(k, i)$ can be considered as unchanged on keyframe candidates when the resolution increases. On the other hand, the content variation metric $D(k, i)$ in equation (3.7) is defined as the difference between $m(k, i)$ in current frame and the mean $\bar{m}(k - 1, i)$ among previous frames. When a keyframe candidate is under evaluation, the content variation $D(k, i)$ turns to be smaller when the resolution increases due to the unchanged $m(k, i)$ and the increased mean $\bar{m}(k - 1, i)$. It also means the threshold can be decreased accordingly to detect the keyframe. However, the standard deviation $\sigma(k - 1, i)$, as a part of the threshold in equation (3.8), increases with larger resolution especially when the sequence contains more motion scenes. Hence, it is reasonable to assume that the function γ is inversely proportional to the resolution so that when resolution is increased the threshold can still be scaled down accordingly even with an increasing standard deviation.

Resolution of a video is usually measured by " $Width \times Height$ " in pixel. It seems difficult to directly incorporate the true resolution into the calculation of dynamic threshold. Thus we first define a 'resolution factor', which is the ratio between a base resolution $base_res$ (176x144) and the resolution of current video. While analyzing data from the training sequences, it was observed that the relationship between γ and the resolution factor is nonlinear and γ is proportional to the logarithm of the resolution factor. Thus, now we have

$$\gamma \propto \log_2 \frac{base_res}{resolution}. \quad (3.9)$$

Correlation Between the QP and the Threshold

Motion estimation is related to the QP used in the sequence. Video will be coded with a higher quality using a smaller QP. So the number of intra-coded macroblock would increase in each frame when the QP is decreased at the same resolution. To show the behavior, the trends on changes of mean $\bar{m}(k, i)$ and standard deviation $\sigma(k, i)$ are plotted in Figure 3.3 for one concatenated sequence when QP setting is changed from 40 to 20 at CIF resolution. Same as Figure 3.2, although the testing has been done on many sequences, only the data obtained from one testing sequence is illustrated here. The sequence consists of 100 frames from "Crew" and 100 frames from "Soccer", and it is coded with IPP...P structure by H.264 encoder [55] as in previous subsection. It is very clear that both $\bar{m}(k, i)$ and $\sigma(k, i)$ are larger with a smaller QP. Following the same considerations as in previous discussion on resolution, the threshold should be relatively decreased when the QP is decreased. It is thus reasonable to assume that γ is proportional to QP. However, due to the continuous characteristic of the QP value, the threshold cannot be exactly adaptive to every small change on QP. Thus for simplicity, the entire QP range is roughly divided into two categories, with $QP \leq 32$ corresponding to high

FIGURE 3.3: (Left) Changes on mean of IMBs; (Right) Changes on standard deviation of IMBs at different QPs

quality videos and $QP > 32$ for low quality videos. So we have

$$\gamma \propto \left\lceil \frac{\overline{QP}}{32} \right\rceil \quad (3.10)$$

where $\lceil x \rceil$ is the ceiling operation that returns the smallest integer value that is greater than or equal to the number x and $\overline{QP}$ denotes the average value of QPs at the frame level.

Dynamic Threshold

Based on the above deduction, the γ function is defined as follows:

$$\gamma = a + b \cdot \log_2 \frac{base_res}{resolution} + c \cdot \left\lceil \frac{\overline{QP}}{32} \right\rceil \quad (3.11)$$

where a , b , and c are three constants. The objective of an optimal algorithm is the correct detection of all keyframe candidates so that these keyframes can be intra-coded, leading to the highest global PSNR gain. In this method, the parameters a , b and c are obtained empirically, and are fine-tuned during the numerical analysis in order to establish the best combination such that the threshold is near optimal for keyframe detection under all kinds of scenarios.

The empirically obtained values are $a = 4$, $b = 0.5$ and $c = 1$. The dynamic threshold for keyframe detection at frame k in the i th GOP is thus proposed to be:

$$T(k, i) = \left\{ 4 + 0.5 \cdot \log_2 \frac{\text{base_res}}{\text{resolution}} + \left\lceil \frac{\overline{QP}}{32} \right\rceil \right\} \cdot \sigma(k-1, i) \quad (3.12)$$

Upper and Lower Bounds of the Threshold

The process of keyframe selection is a sequential process and the dynamic threshold is updated after each frame is encoded. The proposed threshold aims to capture frames with abrupt scene change as well as intensive motion movement. It is necessary to add upper and lower bounds in order to make the final threshold $T_f(k)$ robust.

In some cases, the calculated dynamic threshold in equation (3.8) may exceed the total number of MBs in the frame, denoted as N_{total} . It usually happens in sequences containing several complex motion scenes. The large variation on IMBs may lead to a large standard deviation through the threshold updating process which makes $T(k, i)$ overshoot N_{total} at frame k in the i th GOP. For example, coding a concatenated video starting with sequence "Soccer" at CIF resolution with QP=20 and a scene change happens at frame $k = 100$, the accumulated standard deviation at frame $k = 100$ is $\sigma(99, i) = 100$ and the mean is $\overline{m}(99, i) = 78$. By calculating the threshold using equation (3.12), we get $T(k, i) = 400$ that overshoots $N_{total} = 396$. When this happens, the threshold needs to be refined. It is observed in testing that, usually the mean $\overline{m}(k, i)$ of IMBs is also large when $T(k, i)$ overshoot N_{total} . If a scene change appears at this point, the content variation $D(k, i)$ would end up with a relatively small value and the refined threshold should take this into consideration. Therefore, in this work, an upper bound T_{UP} of the IMBs for keyframe detection is set. The value of T_{UP} has been

selected based on [50] and on statistical empirical analysis, recommending a security value of $T_{UP} = 95\% \times N_{total}$.

For some sequences in which the scenes are relatively static or the motion is slow and smooth, the mean and standard deviation of IMBs would be very small. In this case, even a small content variation detected in a medium motion frame would be considered as sufficient to falsely detect the frame as a keyframe. Taking the same example in the discussion of upper bound, for sequence "Soccer", the motion scene in the first few frames is not complex. At frame $k = 10$, the standard deviation $\sigma(9, i) = 6$ and $\bar{m}(9, i) = 14$ are very small. The number of IMBs $m(10, i) = 92$ in frame 10 and it should not be detected as a keyframe. However, the content variation $D(k, i) = 78$ is still found to be larger than the calculated threshold $T(k, i) = 56$ and this leads to a fault detection at frame 10. To avoid too frequent keyframes, a lower bound T_{LO} of the IMBs is also set. T_{LO} should be much smaller than T_{UP} but it should be large enough to avoid fault detection on small content variations. The value of T_{LO} is experimentally obtained and it is set to $T_{LO} = 50\% \times N_{total}$. With the upper and lower bounds, the refined threshold $T_f(k, i)$ is then defined as:

$$T_f(k, i) = \max \{ (T_{LO} - \bar{m}(k-1, i)), \min \{ T(k, i), (T_{UP} - \bar{m}(k-1, i)) \} \}. \quad (3.13)$$

3.3.3 Keyframe Detection Algorithm

This section outlines the encoding process of a sequence with keyframe detection algorithm. The initial coding structure is set to IPP...P in encoder configuration file. Keyframes will be detected during encoding and they will then be re-encoded as INTRA frames. Since the resolution and QP value are known, we can calculate γ before the encoding starts. The first

frame of a sequence is always intra-coded, so the next frames use this first keyframe as a reference.

Before a new frame to be encoded, the mean m_{k-1} , standard deviation σ_{k-1} and threshold $T_f(k)$ are updated based on the information gathered from previous frames. The new frame k is firstly inter-coded and the counter IMB_k records the total number of intra-coded macroblocks in frame k . For every intra-coded macroblock is recorded, $D(k) = IMB_k - \bar{m}_{k-1}$ is computed and compared with the threshold $T_f(k)$. If $D(k) > T_f(k)$, the inter-coding of the frame is aborted and the frame is re-encoded as an INTRA frame, establishing a new keyframe. Parameters m_k and σ_k are reset to 0. Otherwise, if $D(k) \leq T_f(k)$, the mean $\bar{m}_k$, standard deviation σ_k , and threshold $T_f(k+1)$ are updated and the encoding proceeds to subsequent frame $(k+1)$ while there are frames left to be encoded.

3.4 Rate Control with Adaptive GOP size

In this section, the proposed keyframe detection technique is applied to rate control scheme to determine the sizes of GOPs properly. The size of each GOP is the additional output in the proposed rate control scheme and it is adaptive to the content of the input video sequence.

3.4.1 Two Requirements on the Sizes of GOPs

Let $N(i)$ be the number of pictures in the i th GOP. For a group of GOPs from the 1st to i th group, the average value of $N(j)$ ($j = 1, 2, \dots, i$) is computed as

$$\bar{N}(i) = \frac{1}{i} \sum_{j=1}^i N(j). \quad (3.14)$$

There are two requirements on the value of $N(i)$ and they are stated as below.

1. Since an INTRA frame is a random access point and it is also needed to prevent error propagation, the value of $\bar{N}(i)$ is required not to be greater than a predefined fixed GOP size N_f , i.e.

$$\bar{N}(i) \leq N_f ; \forall i, \quad (3.15)$$

2. Based on [45], the probability of the subsequent scene change within the next few frames after the current scene change is very low. To improve the coding efficiency, a lower bound on the value of $N(i)$ is set. On the other hand, the INTRA frame is a synchronization point in the video stream to limit the error propagation. Too large GOP sizes should be avoided. An upper bound is also set on the value of $N(i)$. In other words, there exist N_L and N_U such that

$$N_L \leq N(i) \leq N_U ; \forall i, \quad (3.16)$$

where $N_L < N_f$ and $N_f < N_U$.

3.4.2 Computation of $N(i)$

For simplicity, the number of frames that is already coded in the i th GOP is denoted as $n(i)$. According to these two requirements on the sizes of GOPs, the upper bound of $N(i)$, $\hat{N}_u(i)$, is computed as

$$\hat{N}_u(i) = \min\{N_U, i \times N_f - \sum_{j=1}^{i-1} N(j)\}. \quad (3.17)$$

Figure 3.4 depicts the flowchart of the proposed rate control scheme on GOP size determination. Abrupt scene change detection is based on the dynamic threshold based keyframe detection algorithm proposed in Section 3.3.

FIGURE 3.4: The diagram of the proposed scheme to determine GOP size

3.4.3 Bit allocation and QP determination

We focus on the case that the GOP structure is IPPP...P and the bit rate is constant. Since the information of exact GOP size is not known in the beginning of the i th GOP, the total number of bits allocated to the i th GOP is computed based on the upper bound of the i th GOP, $\hat{N}_u(i)$, as follows:

$$G(0, i) = \frac{b_r}{F_r} \times \hat{N}_u(i) + \left[\frac{b_r}{F_r} \times \sum_{m=1}^{i-1} N(m) - \sum_{m=1}^{i-1} \sum_{n=1}^{N(m)} f_A(n, m) \right] \quad (3.18)$$

where b_r is the available channel bandwidth, F_r is the frame rate and $\hat{N}_u(0) = N_f$ for the first GOP. $f_A(n, m)$ is the actual bits generated from the n th frame in the m th GOP. It can be seen from (3.18) that the bit allocation of a GOP depends on those of the previous ones. In order to make sure all GOPs end up with an almost uniform quality, each GOP should use its own budget and the buffer occupancy after coding one GOP should be back to the predefined initial buffer fullness B_0 . Thus if the buffer occupancy $B_c(0, i)$ after coding $(i - 1)$ th GOP cannot reach B_0 , the mismatch is also considered in the bit allocation of i th GOP as shown in equation (3.18). In the coding process, the remaining bits at k th frame in i th GOP is updated frame by frame as $G(k, i) = G(k - 1, i) - f_A(k - 1, i)$, ($k = 1, 2, \dots, \hat{N}_u(i) - 1$) where $f_A(k - 1, i)$

is the actual bits of $(k - 1)$ th frame. But if there is a mismatch between the allocated bits and the actual coded bits in the m th GOP, especially when a frame is detected before the GOP size reaches $\hat{N}_u(m)$, the mismatch should be compensated when allocating bit to the $(m + 1)$ th GOP.

Same as [39], the QP of the I frame and the first P frame in the first GOP is predefined according to the values of b_r and $\hat{N}_u(i)$. The I frame and first P frame of other GOPs are using the GOP-level starting QP that is related to the average QP of P frames in the previous GOP, the values of b_r and $\hat{N}_u(i)$. The QPs of subsequent P frames are computed using a quadratic model [37] according to the target bits that are assigned to each P frame. For basic unit layer rate control, the approach in JM [39] is adopted in this work.

3.4.4 Algorithm flowchart

The flowchart in Figure 3.5 outlines the coding process with rate control of a video sequence. The first frame of a sequence is always INTRA-coded as a keyframe. Since the accurate GOP sizes are not known until the sequence is coded and the content variation are measured, the GOP level bit allocation is done based on the calculated upper bound of $N(i)$ for the i th GOP. Before a new frame k is coded, the number of frames already coded in the i th GOP, $n(i)$, is checked according to the two requirements of determining GOP size proposed in Section 3.4.1. If $n(i)$ can meet the two requirements, the new frame k is first INTER-coded with the QP obtained from the frame level bit allocation. The counter $m(k, i)$ records the total number of INTRA MBs in the frame k . Whenever an abrupt scene change is detected on frame k , i.e. $D(k, i) > T_f(k, i)$, the encoding process of frame k as INTER frame stops, and the frame is re-encoded as an INTRA frame that starts a new GOP. The threshold and all the counters are updated accordingly.

FIGURE 3.5: Flowchart of the proposed algorithm.

3.5 Experimental Results

The performance of the proposed algorithm is evaluated in this section with two parts of experiments. Traditional detection and retrieval problems use precision and recall as a basic measure to evaluate the performance. The precision and recall measure the proportion of the detected entities that are properly detected. In this work, the target is not only to detect the frames where the temporal correlation is low, but also to choose better position for the keyframes that leads to a better RD performance.

In the first part, the RD performance of the proposed keyframe detection algorithm is compared with the conventional fixed GOP method and the method in [50]. The proposed dynamic threshold model was implemented on H.264 reference software JM [55] - main profile. Conventional method with fixed GOP size adopts two commonly used configurations, i.e., $GOP = 16$ and $GOP = 32$. The method proposed in [50] has also been implemented on the same platform for performance comparison. Different video sequences are combined to generate test sequences at three different resolutions: QCIF, CIF and 4CIF. The configuration of the test sequences are listed below. All the sequences were coded at a framerate of 15fps in all the methods being compared. For each sequence, 4 QP settings ($QP_P = 28, 32, 36, 40$) were used to code P frames and a smaller QP ($QP_I = QP_P - 2$) was used for I frames accordingly in order to improve the quality of reference frames.

QCIF resolution:

A: Crew (100 frames) + Soccer (100 frames) + Harbor (100 frames);

B: City (120 frames) + Bus (75 frames) + Foreman (90 frames);

C: Ice (90 frames) + Foreman (100 frames) + Akiyo (90 frames).

CIF resolution:

D: Crew (100 frames) + Soccer (100 frames) + Harbor (100 frames);

E: City (120 frames) + Akiyo (120 frames) + Foreman (60 frames).

4CIF resolution:

F: Crew (100 frames) + Soccer (100 frames) + Harbor (100 frames);

G: City (120 frames) + Harbor (100 frames) + Ice (80 frames).

FIGURE 3.6: RD performance comparison on QCIF sequences

Figure 3.6 - Figure 3.8 plot the RD curves of the four methods and the corresponding average PSNR gains achieved by the proposed method are also shown in Table 3.1. Average PSNR gain is calculated via the software AVSNR [56]. It can be seen that the proposed method with dynamic threshold keyframe detection shows better RD performance than both GOP=16 and GOP=32. And the obtained gain is higher when the sequence comes with less motion, for

FIGURE 3.7: RD performance comparison on CIF sequences

FIGURE 3.8: RD performance comparison on 4CIF sequences

example, Sequence E that consists of low motion sequence "Akiyo" and moderate low motion sequence "Foreman", achieves higher gain than Sequence D consisting of relatively high motion sequences "Crew" and "Soccer". This is because in low motion sequences, an INTRA frame is more expensive in terms of bit consumption as compared with an INTER frame if the temporal correlation among consecutive frames is strong. Thus the periodical insertion of INTRA frame as in conventional method results a bigger drop in coding efficiency. The method in [50] also does content-based keyframe insertion and it can be seen that the proposed method attains comparable RD performance as compared with [50]. But the proposed method is much more flexible as it takes resolutions and QPs directly for calculation, whereas the method in [50] is constrained by predefined parameters that are determined by limited

combinations of encoder configuration.

TABLE 3.1: Average PSNR gain for Proposed Dynamic Threshold Based Keyframe Selection

Sequence		PSNR Gain (dB)		
		v.s. GOP16	v.s. GOP32	v.s. [50]
QCIF	A	0.336	0.149	0.046
	B	0.550	0.292	0.050
	C	0.934	0.430	0.056
CIF	D	0.223	0.112	0.030
	E	0.918	0.471	0.021
4CIF	F	0.147	0.060	0.022
	G	0.215	0.072	0.035

It is noted that the proposed algorithm requires some extra computations as compared with the scheme without keyframe insertion, i.e., IPP...P structure. As it can be seen from Section 3.3, the calculation of the dynamic threshold is quite simple. The computation on both mean and standard deviation as well as the update of threshold is only carried out at frame level. The processing time increased from that part can be ignored. The extra processing time is only noticeable when a keyframe is detected and in this case, the keyframe has to be re-encoded as an INTRA frame after it has been partially coded as an INTER frame. Therefore, the overall processing time of a keyframe can be denoted as,

$$t_{KF} = \alpha \cdot t_P + t_I, \quad (3.19)$$

where $\alpha \subseteq [0, 1]$ represents the part of frame that has already been inter-coded before it is detected as keyframe, t_P and t_I denote the time required to code the detected keyframe as INTER and INTRA frame, respectively. A ratio between average t_{KF} and average t_P is then computed, i.e. $R_{KF} = t_{KF}/t_P$, for all the keyframes in the test sequences as shown in Table 3.2. It can be seen that the ratio is almost one in all the tests, which means the detection of keyframe and re-encode the frame as an INTRA frame only requires a little extra computation as compared with conventional method without keyframe detection. The

total encoding time of a sequence with and without applying the proposed method is also compared. The percentage of relative encoding time increment is calculated as

$$\Delta t_{total} = \frac{t_{prop} - t_{conv}}{t_{conv}} \cdot 100\%, \quad (3.20)$$

where t_{prop} is the total encoding time when the proposed method is used and t_{conv} is the total encoding time without any keyframe insertion. Table 3.2 shows that the increased encoding time is no more than 0.1% in all the tests. And for sequences with low correlated high complexity INTER frames, we can see a negative sign there which indicates a decrease in overall processing time. This is because when a keyframe is inserted at the optimal location, the amount of encoding time reduced on subsequent low correlated INTER frame is more than the extra processing time incurred on the keyframe. Therefore, although the detection of keyframe as well as re-encoding the frame as an INTRA frame requires slightly more computation on the frame, the added computational cost is negligible, and sometimes even improves the whole encoding process leading to decreased global processing time. A bigger gain in processing time can be expected if the video is longer and with more candidates for keyframes. The processing time analysis in this section is the same as that proposed in [50] and similar results are obtained.

TABLE 3.2: Processing time for Proposed Dynamic Threshold Based Keyframe Selection

Sequence	R_{KF}					Δt_{total}			
	QP =	28	32	36	40	28	32	36	40
QCIF	A	1.09	1.12	1.09	1.06	0.10%	0.09%	0.07%	0.02%
	B	1.08	1.07	1.10	1.09	0.06%	0.05%	0.07%	0.06%
	C	1.10	1.09	1.09	1.04	0.08%	0.07%	0.07%	0.03%
CIF	D	1.08	1.12	1.10	1.09	-0.02%	-0.02%	-0.01%	-0.01%
	E	1.12	1.10	1.06	1.07	0.08%	0.10%	0.04%	0.05%
4CIF	F	1.05	1.05	1.04	1.03	-0.03%	-0.01%	-0.01%	-0.01%
	G	1.06	1.04	1.04	1.04	0.06%	0.03%	0.02%	0.02%

In the second part, the performance of the proposed algorithm applied in rate control is

evaluated. The algorithm was also implemented on H.264 reference software JM [55] - main profile and the performance is compared with the rate control in [39] with a fixed GOP size as 30. Two types of testing sequences were used in this part. Type 1 sequences were manually generated by concatenating 300 frames from 9 different standard video sequences, namely "Akiyo", "Bus", "Crew", "City", "Foreman", "Harbor", "Ice", "Mobile" and "Soccer". Each standard video sequence was separated into small segments and each segment varies from 25 to 60 frames. Testing sequences at three resolutions (QCIF, CIF and 4CIF) were composed by mixing different segments from different standard sequences. Small segments from each standard sequences were spread into different testing sequences so that the contents in different testing sequences were totally different. Type 2 sequences were natural video clips of news or movie trailers downloaded from YouTube. Each testing sequence was tested on two targeted bit-rates. The lower and upper bounds for GOP size were set as $N_L = 5$, $N_U = 50$ in the experiments.

Table 3.3 shows the performance comparison between the proposed new rate control with adaptive GOP and JM rate control. The average relative rate control error (arRCE) is measured by

$$arRCE = \frac{|A - T|}{T} \cdot 100\% \quad (3.21)$$

where T and A are the target and the actual bits of a sequence respectively. It can be seen that the proposed algorithm achieves higher average PSNRs than the rate control in [39] at similar arRCEs. It is also observed from the experiments that the gains are higher in sequences where abrupt scene change happens more frequently. This also explains where the gains come from. Frames which have low correlation with reference frames are usually INTRA coded and become reference frames for subsequent frames in the proposed method. Temporal redundancy among frames with strong correlation is better removed and it ultimately leads

TABLE 3.3: Performance comparison between proposed rate control and JM rate control

Sequence format		TBR kbps	JM rate control			Proposed			PSNR Gain (dB)
			Actual Bitrate	arRCE (%)	PSNR (dB)	Actual Bitrate	arRCE (%)	PSNR (dB)	
1	QCIF	25	25.04	0.16	28.58	25.03	0.12	28.77	0.19
		80	80.18	0.23	34.24	80.13	0.16	34.45	0.21
2	QCIF	50	52.64	5.28	31.19	50.23	0.46	31.70	0.51
		100	99.99	0.01	35.50	100.3	0.30	36.05	0.30
3	QCIF	35	35.01	0.03	26.32	35.34	0.97	26.83	0.51
		120	120.03	0.03	31.59	120.03	0.03	32.67	1.08
4	QCIF	20	20.25	1.25	30.94	20.34	1.70	31.78	0.51
		60	60.60	1.33	36.21	60.77	1.28	37.27	1.06
5	CIF	90	92.18	2.42	28.99	91.15	1.28	29.25	0.26
		350	350.19	0.05	34.54	350.17	0.05	34.75	0.21
6	CIF	80	80.04	0.05	28.61	79.98	0.03	28.79	0.18
		300	299.87	0.04	34.51	300.08	0.03	34.91	0.40
7	CIF	300	300.03	0.01	35.97	300.70	0.23	36.29	0.31
		500	499.96	0.001	38.23	500.76	0.15	38.75	0.52
8	CIF	100	100.02	0.02	30.48	100.48	0.48	30.68	0.20
		300	299.95	0.01	35.89	300.91	0.30	36.13	0.24
9	4CIF	300	300.32	0.07	29.76	300.37	0.12	29.92	0.16
		1000	1000.2	0.02	34.40	1000.19	0.02	34.48	0.08
10	320 × 128	30	31.98	6.60	32.65	30.92	3.01	33.40	0.75
		80	80.94	1.18	37.50	82.00	2.5	38.18	0.68
11	480 × 256	70	70.03	0.04	32.12	70.35	0.50	32.19	0.07
		190	189.97	0.02	38.03	195.94	3.12	38.28	0.25

to a gain in coding efficiency.

3.6 Summary

In this chapter, a dynamic threshold method for keyframe detection has been proposed to control GOP size. The dynamic threshold is content adaptive and it is related to the resolution, QP as well as the statistical properties of video sequence. It can accurately capture the frames with either abrupt scene change or intensive motion movement and the detected frames will be intra coded as a keyframe in order to improve coding efficiency. Based on the keyframe detection scheme, a new method of properly determining GOP sizes in rate control has also been proposed. The sizes of GOPs in rate control are now adaptive to the content of the input video sequence and a good trade off between coding efficiency and random access are

obtained. The temporal redundancy among different frames are better removed. Experiment results have shown that the proposed method achieves higher average PSNR than existing rate control scheme with a fixed GOP size.

Chapter 4

Implicit Bit Allocation and Cross-Layer Bit Allocation on SVC

4.1 Introduction

In Chapter 3, the bit allocation problem within a video source coded for the requirement of a single user is studied. We now move to the case where a video source is coded to serve multiple users. As reviewed in Chapter 2, the scalable extension of H.264/AVC (SVC) can be used to deliver video to various clients with different decoding, display capabilities and different network bandwidths over heterogeneous networks. This is especially useful for scenarios where the downstream clients' network conditions, system resources and capabilities are not known in advance. The coding efficiency of SVC base layer is almost the same as single-layer coding but the gaps between SVC and single-layer coding are significantly increased in enhancement layers. This is due to the bottom-up encoding process employed in current SVC. An organization may have different preferences on different layers according

to its user composition. If one of the enhancement layers is used by most users in the organization, current SVC couldn't provide optimal coding efficiency to that layer and thus it is not optimized from the commercial point of view. Schemes that can incorporate customer's composition in coding and schemes to improve coding performance of enhancement layers to reduce the gap between SVC and single-layer coding are desired. Since SVC adopts a layered structure, the problems are bit allocation among layers and in each layer, the bit allocation between motion information and residual information is also an important aspect that affects performance.

In this chapter, three schemes are presented. The customer composition adaptive scalability is proposed in Section 4.2 so that the SVC encoder can optimize a specific layer according to customer composition. Usually the layer configurations, or in another word, the bits allocated to each layer are fixed and they are decided by the range of client capabilities. Thus, the implicit bit allocation, which adjusts the bit allocation between motion information and residual information under the fixed available resources is studied in Section 4.2. In Section 4.3, a motion refinement scheme is then introduced to provide finer granularity in the most important scalable bit rate range. Lastly, if the layer configurations are not balanced in bit allocation, to get a balanced SVC, an improved inter-layer prediction scheme is proposed in Section 4.4 for layers with high resolution. Temporal scalability is enabled by hierarchical coding structure of SVC. The coding efficiency of different temporal levels can be easily adjusted by setting quantization parameters. Thus this chapter focuses on combined spatial and SNR scalability.

4.2 Customer Oriented Combined Spatial and SNR Scalability

4.2.1 Motivation

The idea of SVC is to encode a video content once at the the highest required resolution and bitrate but the resulting scalable bitstream can be extracted and decoded partially to serve many users. Besides all the coding mechanisms inherited from H.264/AVC [6], the inter-layer prediction scheme [18] is introduced to remove redundancy between an enhancement layer and its reference. However, due to the bottom-up coding structure adopted in SVC, the coding efficiency of SVC base layer is almost the same as single-layer coding but the gaps between SVC and single-layer coding are significantly increased in enhancement layers. In addition, current SVC design doesn't consider customers' interest. If a specific high resolution enhancement layer is the most important layer used by most people in an organization, current SVC design is not optimal for that organization from a commercial point of view. Another interesting problem that has not been addressed by current SVC is which reference should be chosen for inter-layer prediction when there is both SNR and spatial scalability involved. This choice affects the coding efficiency of enhancement layers especially those higher enhancement layers.

Since an organization may have different preferences on different layers, a weighting factor that indicates the level of preference should be assigned to each layer based on the customer composition. The coding parameters should be adjusted accordingly so that the quality of different layers correspond with the assigned weighting factors. In this subsection, a new performance index is defined and a new type of region of interest (ROI) is formed to achieve customer oriented scalable coding. The target is to optimize the performance of the layer with highest weighting factor and tolerate some losses in the quality of other layers. The selection

of reference layer based on customers' interest in inter-layer prediction is also discussed with the consideration of customers' interest.

4.2.2 A New Type of ROI and A New Performance Index

In conventional ROI image coding, bits are allocated to different regions in the image according to the relative significance and the desired performance index [57]. The concept is different here for the new type of ROI proposed in this section. Let's first define the "whole region" as the overall scalability range from the bitrate and spatial resolution of base layer to the highest bitrate and resolution required by an organization. A "sub-region" refers to a specific choice of bitrate and resolution range. Resource allocated to each sub-region is fixed, only the tradeoff between motion information and residual information in each sub-region can be adjusted. It is also referred to as implicit bit allocation (IBA).

In the new type of ROI, the control policy of IBA in each sub-region needs to be properly determined so that the performance attained can be in line with the weighting factor. The weighting factor assigned to each sub-region and the total number of sub-regions are decided by the customer composition in the organization. Figure 4.1 depicts the difference between conventional ROI and the proposed new ROI in SVC. The new RIOS are assigned with different priorities so a hierarchical structure can be established. The local ROI, which refers to the most important bitrate range in each spatial layer, has lower priority. The global ROI, which gets the highest priority, corresponds to the most important bit rate range at the most important spatial level in the whole region. The ROI highlighted in Figure 4.1(b) is the global ROI. More detailed description on the hierarchical structure will be illustrated later in this section.

FIGURE 4.1: (a) Conventional ROI in image processing (b) New Type of ROI in SVC

For simplicity, the temporal level is assumed to be fixed at l when considering the combined spatial and SNR scalability. Let j ($j = 1, 2, \dots, J$) indexes spatial levels where $j = 1$ refers to the smallest spatial resolution. i ($i = 1, 2, 3$) indexes SNR levels (i.e. bitrate ranges) and in this section three SNR levels are considered on each spatial level. According to [77], a sub-region corresponding to a specific spatial level j , SNR level i and temporal level l can be represented by a Lagrangian factor $\lambda_{l,j,i}$ (i.e., $\lambda_{l,j,i}$ is related to the bitrate $\phi_{l,j,i}$). All the Lagrangian factors in the whole region are predefined and they satisfy the following:

$$\lambda_{l,j,1} \geq \lambda_{l,j,2} \geq \lambda_{l,j,3} ; \forall j, \quad (4.1)$$

$$\lambda_{l,1,1} \geq \lambda_{l,2,1} \geq \dots \geq \lambda_{l,J,1}, \quad (4.2)$$

$$\lambda_{l,1,2} \geq \lambda_{l,2,2} \geq \dots \geq \lambda_{l,J,2}, \quad (4.3)$$

$$\lambda_{l,1,3} \geq \lambda_{l,2,3} \geq \dots \geq \lambda_{l,J,3}. \quad (4.4)$$

The weighting factor corresponding to each sub-region is determined by customer composition. Sub-regions with more demands get higher weighting factors. Let $w_{l,j,1}$, $w_{l,j,2}$ and $w_{l,j,3}$

denote the three sub-regions at j th spatial level, respectively. A performance index is defined as the performance gap between SVC and the state-of-the-art single-layer coding [6]. The optimization problem is formulated as follows:

$$\min_{\lambda_{lo}^{mv}(l,1), \lambda_{hi}^{mv}(l,1), \dots, \lambda_{lo}^{mv}(l,J), \lambda_{hi}^{mv}(l,J)} \sum_{i=1}^3 \sum_{j=1}^J w_{l,j,i}^2 (\hat{f}_{l,j,i}^* - f_{l,j,i}(\lambda_{lo}^{mv}(l,1), \lambda_{hi}^{mv}(l,1), \dots, \lambda_{lo}^{mv}(l,J), \lambda_{hi}^{mv}(l,J), \wp_{l,1,1}, \dots, \wp_{l,J,3}, w_{l,1,1}, \dots, w_{l,J,3}))^2, \quad (4.5)$$

subject to

$$\wp_{l,j,i} = \hat{\wp}_{l,j,i,0}, \quad i = 1, 2, 3, \quad j = 1, 2, \dots, J,$$

where $\hat{f}_{l,j,i}^*$ is the performance (e.g. PSNR) that can be achieved by the state-of-the-art single layer coding [6], and $f_{l,j,i}$ is the performance achieved by SVC with combined spatial and SNR scalability. $\wp_{l,j,i}$ represents the bitrate of the i th subrange on j th spatial level. $\lambda_{lo}^{mv}(l, j)$ and $\lambda_{hi}^{mv}(l, j)$ are Lagrangian multipliers utilized in ME/MC at spatial level j . They are used to determine the control policy of motion data and residual data in each sub-region.

The key idea of the new ROI is that the bitrate allocated to each sub-region $\hat{\wp}_{l,j,i}$ is fixed and the IBA between motion data and residual data can be adjusted. The target of the new ROI is to ensure that the gap between the coding efficiency of SVC and that of the state-of-the-art single layer coding for a sub-region is a monotonically non-increasing function of its weighting factor, i.e.,

$$|\hat{f}_{l,j_1,i_1}^* - f_{l,j_1,i_1}| \begin{cases} \geq |\hat{f}_{l,j_2,i_2}^* - f_{l,j_2,i_2}|, & \text{if } w_{l,j_1,i_1} \leq w_{l,j_2,i_2} \\ < |\hat{f}_{l,j_2,i_2}^* - f_{l,j_2,i_2}|, & \text{otherwise} \end{cases}.$$

Based on the key ideas of the new ROI, it is needed to determine the value of the Lagrangian multiplier in each sub-region and the cross layer ME/MC in all sub-regions. Assume that

resolution $\varphi(l)$ is selected as the most important resolution in the organization according to the customer composition, i.e.,

$$\varphi(l) = \operatorname{argmax}_j \left\{ \sum_{k=1}^3 w_{l,j,k} \right\}. \quad (4.6)$$

Thus ROIs get different priorities as a hierarchical structure:

- local ROI (lower priority) at j th spatial level, corresponds to $\lambda_{opt}(l, j)$; and
- global ROI (highest priority) in the whole region, corresponds to $\lambda_{opt}(l, \varphi(l))$.

In general, two motion vector fields (MVF) are required at the j th spatial level when the SNR scalable range is wide. For example, if $\hat{\phi}_{l,j,3} \geq 3\hat{\phi}_{l,j,1}$, two pairs of Lagrangian multipliers $(\lambda_{lo}^{mv}(l, j), \lambda_{lo}^{qp}(l, j))$ and $(\lambda_{hi}^{mv}(l, j), \lambda_{hi}^{qp}(l, j))$ are used to generate two MVFs respectively. Each pair of Lagrangian multipliers corresponds to two quantization parameters, in which one is used in ME/MC and the other one is used for quantization of the residual. Thus the two pairs of quantization parameters for the two MVFs are $(QP_{lo}^{mv}(l, j), QP_{lo}^{qp}(l, j))$ and $(QP_{hi}^{mv}(l, j), QP_{hi}^{qp}(l, j))$. The auxiliary variable $\lambda_{opt}(l, j)$ is obtained as

$$\lambda_{opt}(l, j) = \begin{cases} \frac{w_{l,j,1}\lambda_{l,j,1} + w_{l,j,2}\lambda_{l,j,2}}{w_{l,j,1} + w_{l,j,2}}; & \text{if } w_{l,j,1} \geq w_{l,j,3} \\ \frac{w_{l,j,2}\lambda_{l,j,2} + w_{l,j,3}\lambda_{l,j,3}}{w_{l,j,2} + w_{l,j,3}}; & \text{otherwise} \end{cases}. \quad (4.7)$$

The values of $\lambda_{lo}^{mv}(l, j)$, $\lambda_{lo}^{qp}(l, j)$, $\lambda_{hi}^{mv}(l, j)$, and $\lambda_{hi}^{qp}(l, j)$ are given by

$$\lambda_{lo}^{mv}(l, j) = \begin{cases} \lambda_{opt}(l, \varphi(l)); & \text{if } j \leq \varphi(l) \\ \lambda_{opt}(l, j); & \text{otherwise} \end{cases}, \quad (4.8)$$

$$\lambda_{lo}^{qp}(l, j) = \frac{w_{l,j,1}\lambda_{l,j,1} + w_{l,j,2}\lambda_{l,j,2}}{w_{l,j,1} + w_{l,j,2}}, \quad (4.9)$$

$$\lambda_{hi}^{mv}(l, j) = \begin{cases} \max\{\lambda_{lo}^{mv}(l, j), \lambda_{l,j,3}\}; & \text{if } w_{l,j,1} > w_{l,j,3} \text{ and } w_{l,\varphi(l),1} > w_{l,\varphi(l),3} \\ \max\{\lambda_{opt}(l, \varphi(l)), \lambda_{opt}(l, j)\}; & \text{otherwise} \end{cases},$$

$$\lambda_{hi}^{qp}(l, j) = \lambda_{hi}^{mv}(l, j). \quad (4.10)$$

If $\hat{\phi}_{l,j,3} \leq 3\hat{\phi}_{l,j,1}$, one MVF by $(\lambda_{lo}^{mv}(l, j), \lambda_{lo}^{qp}(l, j))$ is enough at temporal level l and spatial level j . The value of $\lambda_{opt}(l, j)$ is calculated as

$$\lambda_{opt}(l, j) = \frac{w_{l,j,1}\lambda_{l,j,1} + w_{l,j,2}\lambda_{l,j,2} + w_{l,j,3}\lambda_{l,j,3}}{w_{l,j,1} + w_{l,j,2} + w_{l,j,3}}. \quad (4.11)$$

The values of $\lambda_{lo}^{mv}(l, j)$, $\lambda_{lo}^{qp}(l, j)$, $\lambda_{hi}^{mv}(l, j)$, and $\lambda_{hi}^{qp}(l, j)$ are calculated to be

$$\lambda_{lo}^{mv}(l, j) = \begin{cases} \lambda_{opt}(l, \varphi(l)); & \text{if } j \leq \varphi(l) \\ \lambda_{opt}(l, j); & \text{otherwise} \end{cases}, \quad (4.12)$$

$$\lambda_{lo}^{qp}(l, j) = \lambda_{hi}^{qp}(l, j) = \lambda_{hi}^{mv}(l, j) = \lambda_{lo}^{mv}(l, j). \quad (4.13)$$

4.2.3 Spatial scalability of Motion Information and Residual Information

The Selection of Reference Layer

Suppose there are six layers, (QCIF, Low), (QCIF, Medium), (QCIF, High), (CIF, Low), (CIF, Medium) and (CIF, High). "Low" presents a layer that is coded with the lowest quality, "High" indicates a highest quality and "Medium" has the quality in between. In the current SVC standard, each slice in enhancement layers is allowed to have one reference layer and it is set to the immediate previous layer by default [99]. Figure 4.2 shows the default reference layer selection for inter-layer prediction in current SVC. The arrows in the figure point to the reference layer of each enhancement layer. However, the reference layer of (CIF, Low) should be properly selected according to the customer composition. Considering the following two cases:

FIGURE 4.2: Reference layer selection - Default

Case 1: (CIF, Low) is demanded by many customers. (QCIF, Low) should be chosen as the reference layer for all slices in (CIF, Low), as shown in Figure 4.3(a). The purpose is to allow more "space" so that both motion and texture information can be refined in (CIF, Low). Let us consider an example of combined spatial and SNR scalability for sequence "Foreman". For an enhancement layer that is fixed at (CIF, 192kb/s) (frame rate = 15 f/s), three possible reference layers are available, namely (QCIF, 64kb/s), (QCIF, 128kb/s) and (QCIF, 160kb/s). Different quality combinations for reference layer and enhancement layer are achieved in the

FIGURE 4.3: Reference layer selection - Case 1 and Case 2

three choices. They are: (35.41dB, 35.07dB) for (QCIF, 64kb/s), (39.21dB, 34.03dB) for (QCIF, 128kb/s) and (40.42dB, 32.68dB) for (QCIF, 160kb/s). It can be seen that the enhancement layer attains best performance when the selected reference layer is (QCIF, 64kb/s).

Case 2: (CIF, Low) is demanded by very few customers. As shown in Figure 4.3(b), (CIF, Low) selects (QCIF, Medium) as its reference layer. For (CIF, Low) and (CIF, Medium), the SNR redundancy between them can be removed. Similarly, for (QCIF, Medium) and (CIF, Medium), the spatial redundancy is removed. Thus, (CIF, Medium) and (CIF, High) can attain higher quality at the expense of quality degradation of (CIF, Low).

The Basic ME/MC in the Whole Region.

The ME/MC starts corresponding to parameters $(QP_{lo}^{mv}(l, j), QP_{lo}^{qp}(l, j))$ is called the basic ME/MC in the whole region. Let (dx_0, dy_0) denote an MV in the base layer, which is obtained by the basic ME/MC. A set of MV candidates around (dx_0, dy_0) is defined as

$$\rho(dx_0, dy_0, \delta_x, \delta_y) = \{(dx, dy) | |dx - dx_0| < \delta_x, |dy - dy_0| < \delta_y\}. \quad (4.14)$$

Assuming $F_{2n}(x, y)$ and $F_{2n+1}(x, y)$ are the reference frame and frame to be predicted, respectively. MV (dx_0, dy_0) is coded in the sub-region corresponding to $\lambda_{l,j,1}$. Without loss of generality, it is assumed the sub-region corresponding to $\lambda_{l,j,1}$ is the reference layer of region corresponding to $\lambda_{l,j+1,1}$. The residual information to be coded is $(F_{2n+1}(x, y) - F_{2n}(x -$

$dx_0, y - dy_0$)).

If $(2dx_0, 2dy_0)$ is directly used as MV at the spatial level $(j + 1)$, it may cause problem. Because the MVF at $(j + 1)$ th spatial level may be very different from the upsampled MVF $(2dx_0, 2dy_0)$. If only refining residual information at $(j + 1)$ th spatial level, there would not be enough motion information to support the refined residual information. As a result, it will cause a drop of coding efficiency at high resolution layer. The problem will become even worse if $(j + 1)$ th spatial level is the most important one according to customers' interest. The cross spatial layer ME/MC scheme is thus used to solve this problem by further generating motion information and residual information at $j + 1$ th spatial layer. The detail is given as below.

Cross Spatial Layer ME/MC Scheme

Both motion information and residual information are refined at higher resolution enhancement layer to achieve a good tradeoff between motion and residual information.

Let $Q_{QP_b}()$ represent the quantization operation and the quantization parameter QP_b is decided by $\lambda_{lo}^{qp}(l, j)$. The motion and the residual information at higher resolution corresponding to $\lambda_{lo}^{qp}(l, j + 1)$ are generated by ME/MC between $F_{2n}(x, y)$ and $F_{2n+1}(x, y)$. (dx, dy) is an MV in the set $\rho(2dx_0, 2dy_0, \delta_{xrs}, \delta_{yrs})$ and it is a candidate for refinement. A spatial refinement criterion is defined as

$$D_{sr}(l, 3, 2n, 2n + 1, dx, dy, dx_0, dy_0, \gamma_1, \gamma_2) = \sum_{x,y} \left| F_{2n+1}(x, y) - F_{2n}(x - dx, y - dy) - \frac{\gamma_1 * S_U(IQ_{QP_b}(Q_{QP_b}(D(l, 2, 2n, 2n + 1, dx_0, dy_0))))}{2^{\gamma_2}} \right|, \quad (4.15)$$

where S_U denotes the up-sampling operation, S_D denotes the down-sampling operation, and $D(l, 2, 2n, 2n + 1, dx_0, dy_0) = (S_D(F_{2n+1}(x, y)) - S_D(F_{2n}(x - 2dx_0, y - 2dy_0)))$. All possible values of (γ_1, γ_2) are predefined in the set Ω_{sr} . For example, the values can be $\{(1, 0), (0, 0)\}$. A further ME/MC with an MV (dx, dy) will be carried out at the highest resolution if the

following inequality

$$\begin{aligned} \min_{(\gamma_1, \gamma_2) \in \Omega_{sr}} \{D_{sr}(l, 3, 2n, 2n+1, dx, dy, dx_0, dy_0, \gamma_1, \gamma_2)\} &+ \lambda_{lo}^{mv}(l, 3) \min\{R_{mv}(dx - 2dx_0, dy - 2dy_0), R_{mv}(dx, dy)\} \\ &\leq D_{sr}(l, 3, 2n, 2n+1, 2dx_0, 2dy_0, dx_0, dy_0, 1, 0) \end{aligned} \quad (4.16)$$

is satisfied. If inequality (4.16) cannot be satisfied with any (dx, dy) in the set $\rho(2dx_0, 2dy_0, \delta_{xrs}, \delta_{yrs})$, the remaining energy $D_{sr}(l, 3, 2n, 2n+1, 2dx_0, 2dy_0, dx_0, dy_0, 0, 1)$ is coded directly.

4.2.4 SNR Scalability of Motion Information and Residual Information

For combined SNR and spatial scalability studied in this section, the ME/MC scheme proposed in [63] is adopted here for SNR scalability. The basic ME/MC is started with parameters $(QP_{lo}^{mv}(l, j), QP_{lo}^{qp}(l, j))$ and MV (dx_0, dy_0) is obtained in the sub-region corresponding to $\lambda_{l,j,1}$. If MV (dx_0, dy_0) is used at all SNR layers, the coding efficiency at high bit-rate will be very low. To ensure the performance of the sub-region at high bitrate with the larger weighting factor, a further ME/MC will be performed based on $(\lambda_{hi}^{mv}(l, j), \lambda_{hi}^{qp}(l, j))$ if the inequality

$$\begin{aligned} \min_{(\gamma_1, \gamma_2) \in \Omega_{br}} \{D_{br}(2n, 2n+1, dx, dy, dx_0, dy_0, QP_b, \gamma_1, \gamma_2)\} &+ \lambda_{hi}^{mv}(l, j) \min\{R_{mv}(dx - dx_0, dy - dy_0), R_{mv}(dx, dy)\} \\ &\leq D_{br}(2n, 2n+1, dx_0, dy_0, dx_0, dy_0, QP_b, 1, 0) \end{aligned} \quad (4.17)$$

is satisfied.

Similarly if inequality (4.17) cannot be satisfied with any (dx, dy) in the set $\rho(dx_0, dy_0, \delta_{xrb}, \delta_{yrb})$, then the remaining information $D_{br}(2n, 2n+1, dx_0, dy_0, dx_0, dy_0, QP_b, 1, 0)$ is coded.

4.2.5 Experimental Results

The experiment was carried out using the software package from HHI [68] with $\Omega_{sr} = \Omega_{br} = \{(1, 0), (0, 0)\}$. Four standard video sequences with different characteristics, namely "Foreman", "Bus", "Football" and "Mobile", were tested. Among the four sequences, "Foreman"

contains slow motion activities with a scene change, "Bus" and "Mobile" contain moderate motion activities and "Football" is considered as a high motion sequence with fast object and camera motion. A structure of five layers were used in the experiment and the settings are listed in Table 4.1. The temporal level was fixed at 15fps throughout the whole experiments.

TABLE 4.1: The settings of five layers in experiment 4.2.5

Sequence	Foreman	Football	Bus	Mobile
Layer 0 : (QCIF, Low)	(QCIF, 36kb/s)	(QCIF, 128kb/s)	(QCIF, 96kb/s)	(QCIF, 64kb/s)
Layer 1 : (QCIF, Medium)	(QCIF, 96kb/s)	(QCIF, 300kb/s)	(QCIF, 240kb/s)	(QCIF, 160kb/s)
Layer 2 : (CIF, Low)	(CIF, 144kb/s)	(CIF, 450kb/s)	(CIF, 360kb/s)	(CIF, 256kb/s)
Layer 3 : (CIF, Medium)	(CIF, 300kb/s)	(CIF, 1052kb/s)	(CIF, 960kb/s)	(CIF, 640kb/s)
Layer 4 : (CIF, High)	(CIF, 480kb/s)	(CIF, 1536kb/s)	(CIF, 1440kb/s)	(CIF, 1024kb/s)

In this experiment, it is assumed that QCIF is always more important than CIF, i.e., $\psi(l) =$

1. Based on equations (4.7) - (4.13), the values of $(\lambda_{lo}^{mv}(l,1), \lambda_{lo}^{qp}(l,1))$ was computed as (5.85, 11.18) and $(\lambda_{hi}^{mv}(l,1), \lambda_{hi}^{qp}(l,1))$ was computed as (5.85, 5.85) by the customer oriented scheme. Therefore, two MVFs were generated at QCIF. The MVF at layer 0 was generated using $(QP_{lo}^{mv}(l,1), QP_{lo}^{qp}(l,1)) = (38, 28)$ and MVF at layer 1 was generated with $(QP_{hi}^{mv}(l,1), QP_{hi}^{qp}(l,1)) = (28, 28)$, respectively. We then focused on testing the choice of reference layer for (CIF, Low). Three types of customer compositions were tested.

Type 1: (CIF, Low) is demanded by very few customers. The reference layer of layers 1, 2, 3 and 4 are chosen as Layers 0, 1, 1 and 3, respectively, as shown in Table 4.2. Two MVFs are generated at CIF by using two pairs of quantization parameters as $(QP_{lo}^{mv}(l,2), QP_{lo}^{qp}(l,2)) = (38, 30)$ and $(QP_{hi}^{mv}(l,2), QP_{hi}^{qp}(l,2)) = (30, 30)$. The PSNR of the Y component for all test sequences are shown in Table 4.3.

TABLE 4.2: Reference layer selection in Type 1

Layer ID	Layer 1	Layer 2	Layer 3	Layer 4
Reference layer Selection	Layer 0	Layer 1	Layer 1	Layer 3

TABLE 4.3: Experimental results in Type 1: (CIF, Low) is demanded by very few customers

Sequence	Foreman	Football	Bus	Mobile
Layer 0: (QCIF, Low)	28.78dB	29.10dB	29.52dB	28.56dB
Layer 1: (QCIF, Medium)	37.18dB	34.24dB	35.16dB	33.45dB
Layer 2: (CIF, Low)	29.53dB	30.45dB	26.23dB	23.67dB
Layer 3: (CIF, Medium)	36.56dB	36.47dB	35.45dB	32.86dB
Layer 4: (CIF, High)	38.44dB	38.73dB	38.04dB	35.22dB

Type 2: (CIF, Low) is demanded by many customers, but (CIF, High) is demanded by more customers. The weighting factor of (CIF, High) is thus larger than that of (CIF, Low), i.e. $w_{l,2,3} > w_{l,2,1}$, in this case. Layers 0, 0, 1 and 3 are chosen as reference layers of layers 1, 2, 3 and 4, respectively, as indicated in Table 4.4. Two MVFs are generated at CIF by using $(QP_{lo}^{mv}(l,2), QP_{lo}^{qp}(l,2)) = (38, 30)$ and $(QP_{hi}^{mv}(l,2), QP_{hi}^{qp}(l,2)) = (30, 30)$. Table 4.5 shows the Y component PSNR for all test sequences in Type 2.

TABLE 4.4: Reference layer selection in Type 2 and 3

Layer ID	Layer 1	Layer 2	Layer 3	Layer 4
Base Layer Selection	Layer 0	Layer 0	Layer 1	Layer 3

TABLE 4.5: Experimental results in Type 2: (CIF, Low) is demanded by many customer but $w_{l,2,3} > w_{l,2,1}$

Sequence	Foreman	Football	Bus	Mobile
Layer 0: (QCIF, Low)	28.78dB	29.10dB	29.52dB	28.56dB
Layer 1: (QCIF, Medium)	37.18dB	34.24dB	35.16dB	33.45dB
Layer 2: (CIF, Low)	33.27dB	31.8dB	30.35dB	28.65dB
Layer 3: (CIF, Medium)	35.98dB	35.76dB	35.39dB	32.72dB
Layer 4: (CIF, High)	38.23dB	38.35dB	37.96dB	35.16dB

TABLE 4.6: Experimental results in Type 3: (CIF, Low) is demanded by many customer and $w_{l,2,3} < w_{l,2,1}$

Sequence	Foreman	Football	Bus	Mobile
Layer 0: (QCIF, Low)	28.78dB	29.10dB	29.52dB	28.56dB
Layer 1: (QCIF, Medium)	37.18dB	34.24dB	35.16dB	33.45dB
Layer 2: (CIF, Low)	34.11dB	32.00dB	30.85dB	29.43dB
Layer 3: (CIF, Medium)	35.74dB	35.54dB	35.26dB	32.52dB
Layer 4: (CIF, High)	38.02dB	38.08dB	37.71dB	34.94dB

Type 3: (CIF, Low) is demanded by many customers and $w_{l,2,3} < w_{l,2,1}$. The reference layer

selection would be the same as Type 2 - Layers 0, 0, 1, and 3 are reference layers of layers 1, 2, 3 and 4, respectively. But two MVFs are generated at CIF by using two pairs of parameters as $(QP_{lo}^{mv}(l, 2), QP_{lo}^{qp}(l, 2)) = (38, 38)$ and $(QP_{hi}^{mv}(l, 2), QP_{hi}^{qp}(l, 2)) = (30, 30)$. Table 4.6 shows the Y component PSNR for all test sequences in Type 3.

It can be observed in the experiments that the bitrate configuration for each layer is fixed. In another words, the resources allocated to each layer are fixed in the experiment. When (CIF, Low) is demanded by many customers as shown in Types 2 and 3, (QCIF, Low) is chosen as the reference layer of (CIF, Low) so that more bits are available for refinement of (CIF, Low). With a proper reference layer selection, the quality of (CIF, Low) is improved in Table 4.5 and Table 4.6 as compared with the Type 1 in Table 4.3 where very few customers demand (CIF, Low). In Type 3, when the weighting factor $w_{l,2,1}$ of (CIF, Low) is increased, a proper QP setting for ME/MC is obtained based on equation (4.7)-(4.13) and it can be seen that the quality of (CIF, Low) in Table 4.6 increases accordingly as compared to Type 2 in Table 4.5. At the same time, since the weighting factor $w_{l,2,3}$ of (CIF, High) drops in Type 3, the quality of (CIF, High) also decreased accordingly in Table 4.6 than those in Table 4.5. Apparently, the final coding efficiency of the proposed scheme is adaptive to the customer composition and it can be used to achieve an optimal customer oriented scalable tradeoff in order to maximize the profit. It should be noted that the calculation of proper QP settings for ME/MC increases the complexity. But the calculations involved in equations (4.7) - (4.13) are simple operations and they can be carried out offline before encoding, which means the proposed method does not increase encoding complexity.

4.2.6 Summary

In this work, a new concept of ROI has been proposed, where the method of Lagrangian optimization is employed. The purpose is to adjust the bitrate allocation between motion and residual in order to maximize the overall system performance based on customer composition. The proposed scheme is specially designed for combined SNR and spatial scalability. For a higher layer with better resolution, the selection of reference layer is studied considering the customer composition. The results demonstrate the advantage of the proposed scheme. The superiority becomes more obvious for the type of videos containing high motion in the region of interest.

4.3 Motion Refined Enhanced Coarse Granular Scalability

4.3.1 Motivation

SNR scalability is considered as one of the most important scalability in SVC. Many results have been reported on the performance of SNR scalability in earlier generation scalable codec and the SVC [58]-[61]. These methods have attained comparable good coding performance with that of the state-of-the-art single-layer coding [6] if the scalable bitrate range is narrow, e.g., when the enhancement layer bitrate is about twice of the base layer bitrate. Since only residual information is refined in enhancement layer by these schemes, there will be no accurate motion information available to support residual information in the enhancement layer when the scalable bitrate range is wide. The PSNR gap between the performance of the enhancement layer at high bitrate and that of single-layer coding can be up to 3dB. To overcome this, several cross-layer ME/MC schemes proposed motion refined CGS to fully utilize the features of CGS [63][64]. These motion refined CGS schemes can be used to refine both motion information and residual information at high bitrate. Using these methods, the coding efficiency is improved at high bitrate as compared with the previous SNR schemes [58]-[61].

A new type of ROI has been proposed in previous section, where the whole bitrate region is divided into several sub-regions and each sub-region is assigned a weighting factor according to the customers' interest. The most important sub-region with highest weighting factor is taken as ROI. In this new type of ROI, the bits allocated to each sub-region is fixed but the bit allocation between motion information and residual information can be adjusted so that the performance in each sub-region is ensured in line with its weighting factor. The new ROI is derived as an optimization problem and solved to achieve a globally optimized solution. However, only the tradeoff between motion information and residual information for CGS

was studied in previous section. Another important problem that has yet been addressed is the granularity of the ROI and other subregions where that of the former should be finer than the latter. Therefore, besides CGS, it is also necessary to study a kind of finer granular scalability such that it can be provided in the most important bitrate sub-region and CGS is used in other sub-regions. We call this medium granular scalability. It is studied in this section and a motion refined scheme is also proposed.

It should be noted that, the work presented in this section was completed before the development of Medium-Grain Quality Scalability that was later involved in SVC standard [9]. Although the term used to name the proposed scheme is very similar to the Medium-Grain Quality Scalability in SVC, the two concepts are totally different. Before this work was published, there were only two types of SNR scalability adopted in the standardization of SVC, which were CGS and FGS (as described in Section 2.2.3.3). At that time when the scalable bitrate range between two CGS layers is wide, CGS can only provide coarse scalability by dropping the whole enhancement layer. Any bitrate between two CGS layers can only be reached by FGS which does progressive refinement on residual information only. In aforementioned discussion, it is known that residual refinement results big quality gap to single layer coding if the scalable bitrate range is wide because there is no accurate motion information to support. Therefore, it motivated us to propose medium granular scalability and its motion refined scheme to provide finer scalability as compared to CGS in a wide bitrate range. It is especially useful if certain bitrate range is considered as the ROI in an organization. Providing finer scalability in the ROI is good to maximize the overall profit. (Note that the Medium-Grain Quality Scalability adopted in SVC later could provide similar functionality but with different technique.)

4.3.2 Difference Between Motion Refined CGS and Motion Refined MGS

(Please note that, the short term "MGS" used in the following parts of this section refers to the medium granular scalability proposed in this section. It is not the Medium-Grain Quality Scalability available in current SVC.)

It is well-known that motion refinement scheme is also utilized in the CGS [62]. Essentially, the major difference between the CGS and MGS is that the bitrate range between two CGS layers can be very large while that between two MGS layers is usually small. Other differences between them are summarized in Table 4.7.

TABLE 4.7: Major difference between motion refinement for CGS and MGS

Scheme	Motion refined CGS	Motion refined MGS
Correlation between two adjacent MVFs	can be very weak	strong
Inter-layer residual for prediction between two adjacent layers	adaptive	not all inter-layer modes of CGS are essential for MGS
QP difference in residual information at two adjacent layers	can be very large	around 6

As shown in Table 4.7 that there exists strong correlation between two adjacent MVFs for motion refined MGS. In other words, MVs of each macroblock (MB) in an enhancement layer are either identical to the MB at its base layer, or within a very small range of it. For example, a refinement range in $(-1, +1)$ could be sufficient. It is highly possible that the coding performance can be improved when the pervious coded residual information is used to predict the current residual information. Thus the only difference between the two is coded. However, this is not always true in the motion refined CGS. When the difference between the bitrates of two successive CGS layers is large, the MVFs can be very different. It is very unlikely that

a prediction of the residual information could improve the coding efficiency when the MV is not similar [99].

4.3.3 The Motion Refined MGS Scheme

In this section, only the first two MGS layers are considered and this scheme can be easily extended to other successive layers.

Let $\lambda_{l,j,1}$ and $\lambda_{l,j,2}$ correspond to the first two successive MGS layers at the l th temporal level and the j th spatial level. $F_{2n+1}(x, y)$ is a frame to be predicted, $F_{2n}(x, y)$ is the corresponding reference frame, and (dx, dy) denotes MV. The coding process of the first layer is exactly the same as the non-scalable video coding. The ME/MC starts with the parameter $\lambda_{l,j,1}$ and distortion $D_p(dx, dy)$, i.e. sum of absolute difference (SAD) is calculated as

$$D_p(dx, dy) = \sum_{x,y} |F_{2n+1}(x, y) - F_{2n}(x - dx, y - dy)|. \quad (4.18)$$

The final MV is obtained by using rate distortion optimization to minimize the following performance index:

$$\min\{D_p(dx, dy) + \lambda_{l,j,1}R_{mv}(dx, dy)\}, \quad (4.19)$$

where $R_{mv}(dx, dy)$ is the amount of bits used to encode MV (dx, dy) . Let $Q_{QP_b}()$ be the quantization operation at the first MGS layer, and (dx_0, dy_0) denotes the final MV obtained at the first SNR layer of temporal level l and spatial level j . All motion information and a part of the residual image $Q_{QP_b}(F_{2n+1}(x, y) - F_{2n}(x - dx_0, y - dy_0))$ are encoded and transmitted to the decoder at the first MGS layer. The remaining energy $D_{br}(dx_0, dy_0, dx_0, dy_0, QP_b)$ defined in equation (4.20) should be encoded and transmitted to the decoder at the second MGS layer.

$$\begin{aligned} & D_{br}(dx, dy, dx_0, dy_0, QP_b) \\ = & \sum_{x,y} |F_{2n+1}(x, y) - F_{2n}(x - dx, y - dy) - IQ_{QP_b}(Q_{QP_b}(F_{2n+1}(x, y) - F_{2n}(x - dx_0, y - dy_0)))| \end{aligned} \quad (4.20)$$

Same as CGS, both motion information and residual information can be refined to maximize the coding efficiency at the second MGS layer. A set of candidate MVs around (dx_0, dy_0) is first defined as

$$\rho(dx_0, dy_0, \delta_x, \delta_y) = \{(dx, dy) \mid |dx - dx_0| < \delta_x, |dy - dy_0| < \delta_y\}. \quad (4.21)$$

(dx, dy) is a candidate in the set $\rho(dx_0, dy_0, \delta_{xrb}, \delta_{yrb})$ for the refinement. Let $R_{mv}(dx - dx_0, dy - dy_0)$ represent the amount of bits for coding of the MV difference $(dx - dx_0, dy - dy_0)$. A refined ME/MC will be carried out and the new set of MV (dx, dy) will be obtained for the second MGS layer only if the inequality

$$\begin{aligned} D_{br}(dx, dy, dx_0, dy_0, QP_b) + \lambda_{l,j,2} \min\{R_{mv}(dx - dx_0, dy - dy_0), R_{mv}(dx, dy)\} \\ \leq D_{br}(dx_0, dy_0, dx_0, dy_0, QP_b) \end{aligned} \quad (4.22)$$

is fulfilled. If inequality (4.22) cannot be satisfied with any (dx, dy) in the candidate set $\rho(dx_0, dy_0, \delta_{xrb}, \delta_{yrb})$, the remaining energy $D_{br}(dx_0, dy_0, dx_0, dy_0, QP_b)$ is coded directly at the second MGS layer. As a summary, the two modes in enhancement layer for the motion refined MGS scheme are:

Mode 1: INTRA coding. No motion information will be coded at the second MGS layer, and residual information at the second MGS layer is predicted from the coded residual information at the first one. $D_{br}(dx_0, dy_0, dx_0, dy_0, QP_b)$ is coded directly;

Mode 2: INTER coding. There exists a set of (dx, dy) such that the inequality in (4.22) is satisfied. (dx, dy) refers to refinement of the corresponding MV at the first MGS layer while residual information is predicted from the coded residual information of the first one, i.e., $D_{br}(dx, dy, dx_0, dy_0, QP_b)$ is coded.

4.3.4 Experimental Results

Four standard video sequences, namely "Foreman", "Crew", "Harbor" and "Soccer", with QCIF size at 15fps were tested in the experiments. The experiments attempted to verify the efficiency of the motion refined MGS scheme. First test is the case where overall bitrate range is very narrow on sequence "Foreman". Bitrates for the two layers of test sequence Foreman were set to be 20 kb/s and 40 kb/s, respectively. As each Lagrangian factor λ corresponds to a QP , only the values of QP are provided in Table 4.8. Consider the following six types of configurations for motion information generation.

TABLE 4.8: QPs for the generation of motion information at each layer (Foreman)

MVF Type	1	2	3	4	5	6
QP at layer 1 (20 kb/s)	45	42.5	41	38	45	42.5
QP at layer 2 (40 kb/s)	45	42.5	41	38	39	39.5

Tables 4.9 lists the PSNR of Y component and the actual bitrates which get from the first case. As shown, the gain achieved by the motion refined MGS is not very significant. Therefore, it is not necessary to refine motion information when the bitrate range is narrow, which refers to Mode 1 of the scheme. However, the selection of the QPs to generate MVFs remains crucial. Otherwise, the coding efficiency can be dropped.

Let us now move to the wide bitrate range and consider the case of three MGS layers with bitrates 20kb/s, 40kb/s and 80kb/s, respectively. Four different types of MVFs were tested, and the values of QPs for the generation of motion information at all layers are listed in Tables 4.10.

TABLE 4.9: PSNR (dB) of each type at two layers (Foreman)

MVF Type	1	2	3	4	5	6
PSNR at layer 1 (20 kb/s)	29.04	28.96	28.66	27.79	29.04	28.96
PSNR at layer 2 (40 kb/s)	30.90	31.47	31.66	31.90	31.28	31.53

TABLE 4.10: QPs for the generation of MVF at each layer (Foreman)

MVF Type	1	2	3	4
QP at layer 1 (20 kb/s)	42.5	38.5	33.5	42.5
QP at layer 2 (40 kb/s)	42.5	38.5	33.5	38.5
QP at layer 3 (80 kb/s)	42.5	38.5	33.5	33.5

Note that QPs in Table 4.10 were set such that the MVF Types 1-3 are optimal for Layer 1, Layer 2, and Layer 3, respectively, and Type 4 is optimal for all the layers. The experimental results on the PSNR values is illustrated in Tables 4.11.

TABLE 4.11: PSNR(dB) value at each layer (Foreman)

MVF Type	1	2	3	4
PSNR at layer 1 (20 kb/s)	28.96	28.14	22.53	28.96
PSNR at layer 2 (40 kb/s)	31.47	31.84	31.75	31.51
PSNR at layer 3 (80 kb/s)	33.87	34.88	35.62	34.87

The numbers with bolded font in Table 4.11 indicate that each layer can achieve a better quality when the QP value used to generate MVF is optimal for that layer. Otherwise the

quality may be compromised because there is no accurate motion information available. Take the performance of Layer 3 at Type 1 and Type 4 as an example. It can be observed that PSNR at Type 4 is improved by up to 1 dB compared with that at Type 1. It is because only residual information is refined at Layer 3 for Type 1, while both residual and motion information is refined for Type 4. Type 4 corresponds to Mode 2 of the proposed motion refined MGS scheme where both residual and motion information are refined in enhancement layer when the bitrate changing range is wide. This experiment shows that motion refinement is necessary when the bitrate range is wide.

Finally, all the four sequences when bitrate ranges are wide at four MGS layers were tested. The target bitrate at each layer is given in Table 4.12. The experimental results on the PSNR value and the actual bitrate for all test sequences are also demonstrated in Table 4.12. In this experiment, same QP was used to generate motion information for all MGS layers of the conventional SNR scheme [61] while different QPs were used to generate motion information at each layer in the proposed scheme. The value of QP for the generation of motion information at each layer is listed in Table 4.13. The RD performances are graphically shown in Figure 4.4. It can be observed that motion refined MGS outperforms existing schemes which have no motion refinement at high bitrate. The PSNR gap between these two schemes is up to 0.66dB for video sequences with high motion. The gap enlarges as bitrate range widens. Therefore, the motion refined MGS is very attractive for video sequences with high motion when the bitrate range is wide. It should be acknowledged that the performance gain is attained by sacrificing some computational complexity in motion refinement.

In summary, here are the findings on medium granular scalability.

1. When the bitrate range is narrow, for example, the upper bound is only twice of the lower bound, motion refinement produces a very small gain, if any. It is thus not

TABLE 4.12: Experimental results for MGS

Foreman	target rate(kb/s)		20	40	60	80
	actual rate (kb/s)	existing	19.81	40.21	60.24	79.78
		ours	19.81	40.04	60.45	80.44
	PSNR (dB)	existing	28.96	31.47	32.63	33.33
		ours	28.96	31.50	32.89	33.99
Crew	target rate (kb/s)		56	112	168	224
	Actual Rate (kb/s)	existing	55.69	113.06	167.76	225.50
		ours	55.69	111.50	168.68	222.83
	PSNR (dB)	existing	30.99	33.17	34.30	35.26
		ours	30.99	33.20	34.60	35.60
Harbor	target rate (kb/s)		64	128	192	256
	actual rate (kb/s)	existing	63.40	128.20	191.03	258.20
		ours	63.40	128.84	191.96	258.18
	PSNR (dB)	existing	30.12	32.26	33.44	34.47
		ours	30.12	32.39	33.50	34.45
Soccer	target rate (kb/s)		56	112	168	224
	actual rate (kb/s)	existing	56.51	112.44	167.55	222.08
		ours	56.51	111.61	166.64	223.38
	PSNR (dB)	existing	31.52	33.96	35.32	36.30
		ours	31.52	33.96	35.44	36.62

TABLE 4.13: Values of QPs for the generation of motion information at all layers

Layer		1	2	3	4
Foreman	existing	42.5	42.5	42.5	42.5
	ours	42.5	40	37	33.5
Crew	existing	39.5	39.5	39.5	39.5
	ours	39.5	35.5	32.5	30.5
Harbor	existing	36.5	36.5	36.5	36.5
	ours	36.5	34.5	31.5	29.5
Soccer	existing	38.5	38.5	38.5	38.5
	ours	38.5	34.5	30.5	28.5

FIGURE 4.4: RD performance for Motion Refined MGS

necessary to refine the motion information. However, selection of a proper λ to generate the motion information remains critical.

2. When the bitrate range is wide, the upper bound is more than four times of the lower bound, which warrants a refinement of the motion information, especially for video sequences with high motion.
3. The overall coding efficiency of MGS drops when the number of MGS layers is increased.

4.3.5 Summary

In this section, an MGS scheme has been studied by using the cross SNR Layer ME/MC scheme that is designed for CGS. Both motion information and residual information can be refined in the enhancement layers to obtain a good tradeoff between them at all bitrates. The coding efficiency is improved by up to 1 dB at high bitrate. The coding gain of motion

refinement comes with the cost of extra bits on the prediction residue of MVFs and the inter-layer prediction mode. Compared to CGS, these extra bits are transmitted more frequently in the MGS, which limits the coding gain from motion refinement. But more improvements can be expected if syntax can also be coded more efficiently.

4.4 Balanced Inter-Layer Prediction

4.4.1 Motivation

In each CGS or spatial layer, both the hierarchical motion compensated inter prediction and the intra prediction are implemented as that for single-layer H.264/AVC coding. Removing all kinds of redundancy among layers so that no duplicated information exists in the whole bitstream is an essential step to achieve high coding efficiency. The inter-layer prediction schemes [18][64] are thus designed and incorporated into current SVC [9]. For a slice in the enhancement layer, a corresponding *reference layer* is selected to get rid of the redundancy in motion information and residual information between the reference layer and the enhancement layer. The reference layer can only be a layer with a dependency-id less than the dependency-id of the layer to be predicted. Section 4.2 introduced a combined spatial and SNR scalability scheme which is adaptive to customers' interest. The proper selection of a reference layer for an enhancement layer at which most customers are interested was also discussed. To recall the scenario, let us consider the combined CGS and spatial scalability with four layers, namely (QCIF, R_0), (QCIF, R_1), (CIF, R_2) and (CIF, R_3), where $R_0 < R_1$ and $R_2 < R_3$. (The frame rate is set to 15fps unless otherwise specified throughout this section.) Since only one reference layer can be selected for each slice in an enhancement layer and the reference layer is set to be the immediate previous layer by default [9], a linear nested structure is adopted among all layers. Figure 4.5 illustrates the structure of default reference layer selection in inter-layer prediction. The arrows between two layers in the figure indicate the relationship of enhancement layer and reference layer in the inter-layer prediction.

It can be seen that the inter-layer prediction marked by (1) removes the SNR redundancy between (QCIF, R_0) and (QCIF, R_1). Similarly, the inter-layer prediction marked by (3) removes

FIGURE 4.5: Inter-layer prediction structure by default

FIGURE 4.6: Improved Inter-layer prediction for (CIF,R2)

the SNR redundancy between (CIF, R_1) and (CIF, R_2) . The inter-layer prediction (2) aims to remove the spatial redundancy. However, it was realized in Section 4.2 that a significant coding efficiency drop can be observed in (CIF, R_2) if there is insufficient bits allocated for the combined spatial and SNR refinement from (QCIF, R_1) to (CIF, R_2) . The performance drop at (CIF, R_2) would cause serious problem especially when a lot of customers are interested in that layer. To solve the problem, usage of (QCIF, R_0) as the reference layer for (CIF, R_2) is proposed in Section 4.2. The new inter-layer prediction structure is shown in Figure 4.6. It is shown in Figure 4.7 that the simulation results on the coding efficiency of sequence "Foreman" at Layer 2 (CIF, R_2) improves when its inter-layer prediction reference layer is changed from default Layer 1 (as shown in Figure 4.5) to Layer 0 (as shown in Figure 4.6). This implies that a balanced bit allocation among all layers is very important in SVC for the combined CGS

FIGURE 4.7: Experimental results for Layer 2 - (CIF, R_2) when reference layer of Layer 2 is changed from Layer 1 to Layer 0

FIGURE 4.8: Experimental results for Layer 3 - (CIF, R_3) when reference layer of Layer 2 is changed from Layer 1 to Layer 0

and spatial scalability.

Meanwhile for the same sequence "Foreman", the change in coding efficiency at Layer 3 (CIF, R_3) resulted by the reference layer selection change of Layer 2 are also plotted in Figure 4.8. Surprisingly, the performance of Layer 3 drops when a great improvement on Layer 2 is achieved by changing its reference layer from Layer 1 to Layer 0. This is because in current SVC only one reference layer can be selected for each enhancement layer, the SNR redundancy

between Layer 3 and Layer 2, and the spatial redundancy between Layer 3 and Layer 1 cannot be simultaneously removed at Layer 3 which consequently degrades the overall performance of SVC.

Based on the observation obtained in the above simulation, two techniques are proposed in this subsection to improve the performance of the existing inter-layer prediction with one reference layer for the combined CGS and spatial scalability. First technique is the inter-layer prediction with two reference layers that removes the inherited constraint in current SVC so that both spatial and SNR redundancy can be removed at the same time. Adaptive inter-layer prediction modes are made available in the proposed technique for (CIF, R_3) . A macroblock (MB) at (CIF, R_3) can be predicted from the corresponding block in either (QCIF, R_1) or (CIF, R_2) . With the first proposed technique, the coding efficiency of (CIF, R_3) can be improved, but there is another problem yet to be solved. If there are only customers interested in (QCIF, R_1) and (CIF, R_2) , it is worth considering how to generate the bitstream such that the performance of both layers can be kept. Thus a new concept of auxiliary layer which aims to further improve the balance among all layers is proposed. The two techniques can either be used jointly or separately.

4.4.2 Inter-Layer Prediction with Two Reference Layers

Consider the same example given with four layers in the Section 4.4.1 for simplicity. Suppose that (QCIF, R_0) is selected to be the reference layer of (CIF, R_2) so that the coding efficiency of (CIF, R_2) can be optimized. We now consider coding (CIF, R_3) by using two reference layers: (QCIF, R_1) and (CIF, R_2) , and the inter-layer prediction structure would be the one shown in Figure 4.9.

FIGURE 4.9: The proposed inter-layer prediction structure with two reference layers

Let $F_{2n}(x, y)$ be the reference frame and $F_{2n+1}(x, y)$ be the frame to be predicted at CIF resolution. The corresponding frames at QCIF resolution are $S_D(F_{2n}(x, y))$ and $S_D(F_{2n+1}(x, y))$ where S_D is a down-sampling operation. The corresponding up-sampling operation is denoted as S_U . $\Psi_{QP} = IQ_{QP}(Q_{QP}(\cdot))$ with Q_{QP} as a quantization operation and IQ_{QP} as the corresponding inverse quantization operation. The corresponding residual at QCIF resolution for picture $F_{2n+1}(x, y)$ is denoted as $D(2n, 2n + 1, dx, dy)$ and it is defined by:

$$D(2n, 2n + 1, dx, dy) = S_D(F_{2n+1}(x, y)) - S_D(F_{2n}(x - dx, y - dy)) \quad (4.23)$$

where (dx, dy) represents the motion vector fields (MVF) at a specific layer.

The MVFs at (QCIF, R_0), (QCIF, R_1), and (CIF, R_2) are represented as (dx_0, dy_0) , (dx_1, dy_1) and (dx_2, dy_2) . QP_0 , QP_1 and QP_2 denote the quantization parameters used at the three layers, respectively. The residual at (QCIF, R_0) after motion estimation and motion compensation is $D(2n, 2n + 1, dx_0, dy_0)$, and the reconstructed residual D'_{QL} is

$$D'_{QL} = \Psi_{QP_0}[D(2n, 2n + 1, dx_0, dy_0)]. \quad (4.24)$$

The inter-layer prediction between (QCIF, R_1) and (QCIF, R_0) denoted as arrow (1) in Figure 4.9 removes the SNR redundancy. The prediction can be represented as $D(2n, 2n + 1, dx_1, dy_1) - \gamma_1 * D'_{QL}$, where $\gamma_1 \in \{0, 1\}$ indicates the adaptive *residual_prediction_flag*. Thus the reconstructed (QCIF, R_1) residual is denoted as:

$$D'_{QH} = \Psi_{QP_1} [D(2n, 2n + 1, dx_1, dy_1) - \gamma_1 * D'_{QL}] + \gamma_1 * D'_{QL}. \quad (4.25)$$

The inter-layer prediction between (CIF, R_2) and (QCIF, R_0) which is numbered as arrow (2) in Figure 4.9 removes the spatial redundancy. The process can be seen as $F_{2n+1}(x, y) - F_{2n}(x - dx_2, y - dy_2) - \gamma_2 * S_U(D'_{QL})$, where $\gamma_2 \in \{0, 1\}$. The reconstructed residual at (CIF, R_2) is:

$$D'_{CL} = \Psi_{QP_2} [F_{2n+1}(x, y) - F_{2n}(x - dx_2, y - dy_2) - \gamma_2 * S_U(D'_{QL})] + \gamma_2 * S_U(D'_{QL}). \quad (4.26)$$

For (CIF, R_3), in order to remove SNR and spatial redundancy simultaneously, both (QCIF, R_1) and (CIF, R_2) can be used as reference layers for inter-layer prediction. A combined CGS and spatial scalability refinement criterion for (CIF, R_3) can then be defined as

$$D_{sbr}(2n, 2n + 1, dx, dy, dx_0, dy_0, dx_1, dy_1, dx_2, dy_2, \gamma_3) = \sum_{x,y} |F_{2n+1}(x, y) - F_{2n}(x - dx, y - dy) - \gamma_3 * S_U(D'_{QH}) - \gamma_4 * D'_{CL} + \gamma_1 \gamma_2 \gamma_3 \gamma_4 * S_U(D'_{QL})| \quad (4.27)$$

where $\gamma_3 \in \{0, 1\}$, $\gamma_4 \in \{0, 1\}$. The MVF (dx, dy) at (CIF, R_3) will be obtained by rate-distortion optimization (RDO) process in each macroblock by minimizing the cost function:

$$J = D_{sbr}(2n, 2n + 1, dx, dy, dx_0, dy_0, dx_1, dy_1, dx_2, dy_2, \gamma_3) + \lambda \cdot R_{mv}(dx, dy) \quad (4.28)$$

where $R_{mv}(dx, dy)$ is the amount of bits to encode the MV (dx, dy) .

It can be seen from the equation (4.27) that there are two reference layers - (QCIF, R_1) and (CIF, R_2), and adaptive modes of inter-layer prediction carried out for each slice at (CIF, R_3).

The two modes are:

Mode 1: Predict from (QCIF, R_1), i.e. ($\gamma_3 = 1, \gamma_4 = 0$);

Mode 2: Predict from (CIF, R_2), i.e. ($\gamma_3 = 0, \gamma_4 = 1$);

Therefore, both the spatial redundancy between (QCIF, R_1) and (CIF, R_3) as well as the SNR redundancy between (CIF, R_2) and (CIF, R_3) can be removed at the same time. The coding efficiency of (CIF, R_2) can then be improved without any performance drop at (CIF, R_3) by using the proposed inter-layer prediction with two reference layers.

4.4.2.1 Experimental Results

TABLE 4.14: QP settings for each sequence

Sequence	QP	Layer 0	Layer 1	Layer 2	Layer 3
Foreman	Group 1	42	28	42	28
	Group 2	42	28	40	33
	Group 3	42	28	42	35
	Group 4	42	28	44	37
Crew	Group 1	37	31	44	40
	Group 2	37	31	42	38
	Group 3	37	31	38	34
	Group 4	37	31	37	31
City	Group 1	32	26	44	28
	Group 2	32	26	41	37
	Group 3	32	26	37	33
	Group 4	32	26	32	26
Soccer	Group 1	36	30	44	40
	Group 2	36	30	40	37
	Group 3	36	30	38	35
	Group 4	36	30	36	30

The proposed prediction structure shown in Figure 4.9 is compared with the existing non-adaptive prediction structure illustrated in Figure 4.6. Experiments were carried out on JSVM

[98] with four video sequences - Foreman, City, Crew and Soccer. For each sequence, four groups of QP settings for Layer 0-3 were used to generate the experimental results for the two structures. The detailed QP settings for each layers of the test sequences are listed in Table 4.14.

FIGURE 4.10: Comparison between proposed scheme and JSVM non-adaptive scheme for the sequence "Foreman", "Crew", "City", "Soccer".

The rate-distortion (RD) performances for all the sequences are shown in Figure 4.10. Since the RD performances for layer 0 (QCIF, R_0) to layer 2 (CIF, R_2) are identical for both schemes, only the curves on average Y-PSNR of layer 3 (CIF, R_3) are shown for both schemes in the figures. It is shown in Figure 4.10 that the average PSNR of Layer 3 can be improved by more than 0.2 dB in most sequences when MB level adaptive reference layer selection is enabled.

4.4.3 A New Concept of Auxiliary Layers

In this subsection, a new concept of auxiliary layer is introduced to improve the coding efficiency of layers when the bit allocations are not balanced.

To get a criterion of balanced bit allocation in the case of combined spatial and SNR scalability, some simulations were carried out to get the ratio between the bitrate of high resolution enhancement layer and the bitrate of low resolution reference layer when the two layers are coded at a same QP and the resolution increment is dyadic. Two groups of simulations, one to check the bitrate increment for QCIF to CIF and the other one for CIF to 4CIF, were conducted and the results are shown in Table 4.15. It can be seen that the average increment factor is approximately 3.5.

Therefore, for the four layers (QCIF, R_0), (QCIF, R_1), (CIF, R_2) and (CIF, R_3) where $R_0 < R_1$ and $R_2 < R_3$, if $R_2/R_0 \geq 3.5$, the bit allocation is considered as balanced because sufficient bits are allocated to (CIF, R_2) for the combined spatial and SNR scalability. The performance of (CIF, R_2) and (CIF, R_3) can be maintained by using the proposed inter-layer prediction in Section 4.4.2. However, if $R_2/R_0 < 3.5$, the bit allocation is not balanced and the performance of (CIF, R_2) and (CIF, R_3) would be very low even with the improved inter-layer prediction proposed earlier. Thus, a new concept of auxiliary layer, which is an additional layer, is introduced to be incorporated to inter-layer prediction. In this case, two coding schemes with and without auxiliary layer are compared as below:

Scheme 1: Without Auxiliary Layer.

The coding structure is shown in Figure 4.9. As proposed in Section 4.4.2, (QCIF, R_0) is the reference layer of both (QCIF, R_1) and (CIF, R_2). Both (QCIF, R_1) and (CIF, R_2) are the reference layers of (CIF, R_3). The bitstream is made up by components of BL, EL1, EL2 and EL3, and

TABLE 4.15: Bitrate increment factor for QCIF to CIF and CIF to 4CIF

Resolution	Sequence	QP	BR Increment factor
QCIF to CIF	Bus	QP=28	3.762
		QP=30	3.716
		QP=32	3.688
		QP=36	3.624
	City	QP=28	4.044
		QP=30	4.043
		QP=32	4.042
		QP=36	3.999
	Crew	QP=28	2.905
		QP=30	2.858
		QP=32	2.833
		QP=36	2.799
	Foreman	QP=28	3.259
		QP=30	3.132
		QP=32	3.044
		QP=36	2.913
	Soccer	QP=28	3.968
		QP=30	3.838
		QP=32	3.728
		QP=36	3.376
CIF to 4CIF	City	QP=28	5.382
		QP=30	4.936
		QP=32	4.614
		QP=36	4.324
	Crew	QP=28	3.464
		QP=30	3.340
		QP=32	3.256
		QP=36	3.210
	Soccer	QP=28	3.032
		QP=30	3.015
		QP=32	3.014
		QP=36	3.012
Average			3.571

they are given as follows:

- **BL:** (QCIF, R_0);
- **EL1:** SNR refinement from (QCIF, R_0) to (QCIF, R_1);
- **EL2:** Spatial refinement from (QCIF, R_0) to (CIF, R_2);
- **EL3:** Combined refinement from (QCIF, R_1) and (CIF, R_2) to (CIF, R_3).

The sizes of BL, EL1, EL2 and EL3 are R_0 , $(R_1 - R_0)$, $(R_2 - R_0)$ and $(R_3 - R_1 - R_2 + R_0)$, respectively.

FIGURE 4.11: The inter-layer prediction structure with Auxiliary Layer when $R_2/R_0 < 3.5$.

Scheme 2: With Auxiliary Layer (QCIF, R_A).

(QCIF, R_A), (QCIF, R_0), (QCIF, R_1), (CIF, R_2) and (CIF, R_3), with the structure shown in Figure 4.11, are generated where $R_A \leq R_0$. In this structure, (QCIF, R_A) is the reference layer of both (QCIF, R_0) and (CIF, R_2), (QCIF, R_0) is the reference layer of (QCIF, R_1), both (QCIF, R_1) and (CIF, R_2) are the reference layers of (CIF, R_3). The bitstream comprises BL, AL, EL1, EL2 and EL3, as provided below:

- **BL:** (QCIF, R_A);
- **AL:** SNR refinement from (QCIF, R_A) to (QCIF, R_0);
- **EL1:** SNR refinement from (QCIF, R_0) to (QCIF, R_1);
- **EL2:** Spatial refinement from (QCIF, R_A) to (CIF, R_2);
- **EL3:** Combined refinement from (QCIF, R_1) and (CIF, R_2) to (CIF, R_3).

The sizes of BL, AL, EL1, EL2 and EL3 are R_A , $(R_0 - R_A)$, $(R_1 - R_0)$, $(R_2 - R_A)$ and $(R_3 - R_1 - R_2 + R_A)$, respectively. Apparently, the size of EL2 is enlarged from $(R_2 - R_0)$ to $(R_2 - R_A)$, and this will improve the quality of layer 2 (CIF, R_2).

TABLE 4.16: Experimental results with $R_A = 64\text{kb/s}$

	Target rate (kb/s)		$R_0 = 192$	$R_2 = 256$
Bus	actual rate (kb/s)	without AL	190.38	254.07
		with AL	193.08	258.48
	PSNR (dB)	without AL	34.96	27.39
		with AL	34.64	29.57
		Gain	-0.32	2.18
City	actual rate (kb/s)	without AL	191.85	253.85
		with AL	192.71	254.92
	PSNR (dB)	without AL	41.81	32.34
		with AL	41.41	36.23
		Gain	-0.40	3.89
Crew	actual rate (kb/s)	without AL	192.48	254.80
		AL(64kb/s)	193.26	258.12
	PSNR (dB)	without AL	36.68	33.29
		with AL	36.34	33.64
		Gain	-0.34	0.35
Foreman	actual rate (kb/s)	without AL	192.48	255.98
		with AL	193.19	255.48
	PSNR (dB)	without AL	41.42	34.59
		with AL	41.22	36.45
		Gain	-0.20	1.86
Soccer	actual rate (kb/s)	without AL	191.25	256.47
		AL (64kb/s)	192.18	255.43
	PSNR (dB)	without AL	37.87	30.65
		with AL	37.58	31.70
		Gain	-0.29	1.05

4.4.3.1 Experimental Results on Auxiliary Layers

To demonstrate the efficiency of auxiliary layer, the coding efficiency of Layer 0 (QCIF, R_0) and Layer 2 (CIF, R_2) in the structure without auxiliary layer as shown in Figure 4.9 and the one with auxiliary layer shown in Figure 4.11 are compared. R_0 is set to 192kb/s and R_2 is 256 kb/s so that the bitrate allocation is considered as unbalanced. Two configurations of the new auxiliary layer (QCIF, R_A) were tested with the values of R_A fixed at 64kb/s and 96kb/s, respectively. Although the auxiliary layer makes the rate-distortion performance of (QCIF, R_0) drop, the QP used in the ME/MD of (QCIF, R_A) was properly set so that the quality drop at (QCIF, 192kb/s) would be as little as possible. It can be seen from both Tables 4.16 and 4.17 that the coding efficiency of (CIF, 256kb/s) is greatly improved with a very small loss at (QCIF, 192kb/s). Figure 4.12 and 4.13 shows a visual comparison between the

TABLE 4.17: Experimental results with $R_A = 96\text{kb/s}$

	Target rate (kb/s)		$R_0 = 192$	$R_2 = 256$
Bus	actual rate (kb/s)	without AL	190.38	254.07
		with AL	193.28	256.35
	PSNR (dB)	without AL	34.96	27.39
		with AL	34.59	29.16
		Gain	-0.37	1.77
City	actual rate (kb/s)	without AL	191.85	253.85
		with AL	190.49	257.22
	PSNR (dB)	without AL	41.81	32.34
		with AL	41.29	35.64
		Gain	-0.52	3.30
Crew	actual rate (kb/s)	without AL	192.48	254.80
		with AL	192.13	254.00
	PSNR (dB)	without AL	36.68	33.29
		with AL	36.21	33.54
		Gain	-0.47	0.25
Foreman	actual rate (kb/s)	without AL	192.48	255.98
		with AL	193.26	256.92
	PSNR (dB)	without AL	41.42	34.59
		with AL	41.09	36.14
		Gain	-0.33	1.55
Soccer	actual rate (kb/s)	without AL	191.25	256.47
		with AL	192.72	255.37
	PSNR (dB)	without AL	37.87	30.65
		with AL	37.53	31.48
		Gain	-0.34	0.83

existing scheme without auxiliary layer and the proposed scheme with auxiliary layer on the snapshots of sequence "Foreman" at Frame #35 and "Bus" at Frame #42. It can be shown that auxiliary layer brings a clear visual quality improvement. Thus, the auxiliary layer is very necessary in the case of $R_2/R_0 < 3.5$ and it also proves that having balanced bit allocation among all layers is important for the combined spatial and SNR scalability. Although the two proposed techniques come with an increase of computational complexity, the complexity increment is an one-time increment at encoder side because of the bitstream level scalability feature of SVC that only needs to encode video sources once. Thus the complexity increment at encoder side would have negligible impact to streaming or broadcasting applications, in which encoding can be done offline.

FIGURE 4.12: : Visual quality comparison between (a) existing scheme without AL and (b) the proposed scheme with AL for the sequence "Foreman" Frame #35.

FIGURE 4.13: Visual quality comparison between (a) existing scheme without AL and (b) the proposed scheme with AL for the sequence "Bus" Frame #42.

4.4.4 Summary

In this section, an improved inter-layer prediction has been proposed to enhance the coding efficiency of layers at high resolution when both CGS and spatial scalability are involved in the system. Two reference layers are enabled in the proposed scheme so that both spatial redundancy and SNR redundancy between layers can be simultaneously removed. A new

concept of auxiliary layer that works with the proposed inter-layer prediction with two reference layers has been introduced to relieve the unbalanced bit allocation problem in the system so that the overall performance is significantly improved.

Chapter 5

Joint Bit Allocation for Multi-Program Video Broadcast

5.1 Introduction

As stated in Chapter 1, in typical broadcast applications, such as digital TV broadcast and video surveillance, different video sequences are encoded in parallel and transmitted simultaneously over a bandwidth-limited channel. Coding of each video sequence independently at an equal proportion of the channel bandwidth is a straightforward method to allocate channel bandwidth but the quality fluctuation may be tremendous with time in a particular video and between videos. Rate allocation among multiple videos with a constrained bandwidth becomes an important issue to achieve optimized consistent video quality in the distributive application. Statistical multiplexing is a way to distribute the channel bandwidth dynamically over time, depending on the relative complexity of each video sequence. The goal is to maximize the overall video broadcast quality under bandwidth constraints while attaining

comparable qualities on all received videos (i.e., inter-program fairness), and stable quality in each video (i.e., intra-program smoothness) [69], [79] - [83], [85].

In this chapter, the problem of joint rate allocation for multiple videos is first addressed based on a platform employing scalable video encoder in Section 5.2. Most of the existing approaches are based on non-scalable video encoders, where computationally intensive re-encoding or transcoding is required to control the bitrate of each video. A new statistical multiplexing system is developed by applying the scalable video coding technique to compress the video programs. The rate allocated to each video can then be easily controlled by simple bitstream truncation. In the proposed system, each video is separated into groups of pictures (GOPs) with fixed lengths and the rate allocation is updated at the GOP boundaries. The proposed multiplexing system has contributions in two modules: look-ahead preprocessor and joint rate allocator. In look-ahead preprocessing, a new complexity measure is developed to determine the base layer coding rate of each video program. The new complexity measure takes into account both spatial complexity and temporal complexity of a video sequence. A similarity index (SMI), which is a variation of the Structure Similarity Index (SSIM) [87] is utilized to measure the temporal complexity. As inherited from SSIM, the SMI incorporates Human Visual System (HVS) characteristics but it is modified to match the rate-distortion relationship in video coding. After each video source is compressed into a base layer and several fine granularity scalability (FGS) layers, a joint bit allocator collects the coding statistics generated from the encoders and a piecewise linear model is used to estimate the rate-distortion relation in FGS layers. Bit allocation of each video is determined by solving an optimization problem that aims to achieve a balance between inter-program fairness and intra-program smoothness while keeping the total bitrates of multiple programs conformant to the channel bandwidth. The scheme was implemented to verify its performance, and it

is shown that the proposed scheme obtains better intra-program smoothness than [85] while keeping a similar quality difference across all videos.

With recent advances in display and camera technologies, 3D video are becoming popular for next generation broadcast and streaming applications. We thus study the bit allocation problem among multiple 3D videos later in Section 5.3 of this chapter. Although there are various representation formats available for 3D video, stereoscopic seems to be the most mature technology at this moment [100]. Different coding formats have been developed for stereo videos but there is no standardized format for storage and transmission system. A scalable stereoscopic video coding scheme that inherits some of the features from scalable video coding is first proposed. For the two views in stereo video, the reference view is independently coded to provide backward compatibility while the auxiliary view is coded using the reference view as base layer to remove inter-view redundancy. The scalable format makes the rate control of each video as easy as bitstream extraction. When multiple stereo videos are available in broadcast applications, the bit allocation problem is more complicated as it includes both bit allocation between videos as well as between views. It is well known that human visual system can compensate and conceal the degraded fidelity in one of the two views provided the other view is encoded at a high quality [101]. A joint bit allocation scheme is thus developed to do adaptive bit allocation between views to obtain high compression efficiency and high perceptual quality.

5.2 Joint Rate Allocation For Statistical Multiplexing in Video Broadcast Applications

5.2.1 Background and Related Work

An accurate estimation of each video's complexity model is a prerequisite for efficient rate allocation. Algorithms used to derive video complexity can be broadly classified into feedback approaches and look-ahead approaches. In the feedback approaches [69]-[71], parameters obtained after the encoding of one or several frames are used to predict the complexity model of subsequent frames. These approaches assume the neighboring frames share similar characteristics but often suffer from performance degradation at scene changes. Look-ahead approaches [72], [73], [85] typically employ a pre-processing step to estimate video complexity before encoding. These approaches are more computationally intensive but have wider choices of statistics and can thus obtain a more accurate estimation. Feedback and look-ahead approaches can also be jointly applied [74]-[76] to predict the bit allocation for each of the videos.

To achieve a good statistical multiplexing performance, not only the channel bandwidth needs to be efficiently allocated between video programs, but also the video encoder needs to be accurately controlled. A number of statistical multiplexing algorithms [69]-[76] employ single layer encoders, such as the non-scalable mode of MPEG-2 [3] and H.264/AVC [6], to perform joint rate control. On these platforms, multiple videos have to be re-encoded or transcoded in order to fit into the channel. Other works [83]-[85] exploit the Scalable Video Coding (SVC) extension of H.264/AVC [9] to reduce computational complexity. But as a trade-off, the rate-distortion performance of SVC is usually not as competitive as non-scalable platform [8]. In general, videos encoded by SVC-based platforms are comprised of one base layer and several

enhancement layers. The target bitrate allocated to each encoded bitstream can be easily achieved through simple truncation instead of re-encoding. This makes the control of video encoder easier and more accurate in SVC based platforms. Therefore, efficiently allocating the channel bitrate according to some optimality criterion becomes more crucial in SVC based platforms. Jacobs *et al.* [84] proposed to distribute channel bandwidth proportionately among video sequences according to their relative complexities. But the bandwidth distribution for base layer of each video uses equal allocation. Wang *et al.* pointed out in [85] that, base layer qualities affect the overall performance and the equal distribution of base layer bandwidth is not efficient. They also showed that the rate allocation scheme proposed in [84] results in a large quality difference between different videos. In [85], both the base layer bandwidth and the channel bandwidth are dynamically distributed among video sequences in order to minimize the quality differences across videos. Although the inter-program fairness was taken care in [85], the performance of base layer rate allocation was not evaluated. More importantly, the intra-program smoothness criterion was not considered in their optimization problem, which may cause large quality fluctuations within each video.

5.2.2 System Overview

Figure 5.1 shows the basic structure of the proposed statistical multiplexing system. The system consists of several complexity analyzers, scalable video coding (SVC) encoders, bitstream extractors, a 2-D joint rate allocator and a multiplexer. Each complexity analyzer is used as a preprocessor to analyze a video source on a GOP basis. It evaluates the similarities between neighboring pixels as well as similarities between adjacent frames to derive picture statistics that reflect the complexity of each video source. The 2-D joint rate allocator collects the picture statistics from every analyzer and calculates the base layer bitrate allocation for each

encoder based on the relative complexities of the videos. After receiving the video and base layer bitrate allocation, each SVC encoder encodes the video into a base layer and several FGS layers. The coding statistics are thus generated as byproduct from encoding process. The joint rate allocator takes these coding statistics as input and calculates the optimal bitrate allocation to different video programs in a GOP. The compressed bitstream is then fed into a bitstream extractor which extracts a substream out of the input bitstream according to the target bitrate given by the joint rate allocator. In the last module - multiplexer, the extracted substreams of the different video sources are multiplexed and transmitted over a single channel.

FIGURE 5.1: Structure of the proposed multiprogram video coding system

5.2.3 Similarity Based Complexity Analyzer

The similarity based complexity analyzer as shown in Figure 5.1 is a preprocessor to evaluate the complexity of each video and the obtained complexity is then used for determining the base layer coding rate. It can be observed large content variation between a frame and its reference frame, and large variation of pixel values on a frame are the causes of high complexities in video. Thus the proposed scheme uses both temporal complexity and spatial complexity to characterize the video complexity.

5.2.3.1 Existing Structure Similarity Index

The temporal complexity measures the video complexity on temporal direction, which is related to similarity between adjacent frames. Zhou *et al.* proposed the Structural Similarity Index (SSIM) in [87] as a quality assessment method for measuring similarity between two static images. It considers the characteristics of Human Visual System (HVS) in an attempt to incorporate perceptual quality measures [87]. Essentially SSIM measures the deviations in luminance, contrast and structure between the two images and these deviations are represented as:

$$\begin{aligned} l(m, n) &= \frac{2\mu_m\mu_n}{\mu_m^2 + \mu_n^2} \\ c(m, n) &= \frac{2\sigma_m\sigma_n}{\sigma_m^2 + \sigma_n^2} \\ s(m, n) &= \frac{\sigma_{mn}}{\sigma_m\sigma_n} \end{aligned} \quad (5.1)$$

SSIM as shown below is a product of these three elements.

$$SSIM = \frac{(2\mu_m\mu_n + c_1)(\sigma_{mn} + c_2)}{(\mu_m^2 + \mu_n^2 + c_1)(\sigma_m^2 + \sigma_n^2 + c_2)} \quad (5.2)$$

where m and n represent the two images under comparison; μ_m , μ_n and σ_m^2 , σ_n^2 are the means and variances of two images respectively; σ_{mn} is the covariance of image m and n . c_1 , c_2 are two constants that stabilize the division with weak denominator.

The application of SSIM in a video coding context is still under investigation [91][92]. The conventional usage of SSIM as quality assessment metric has been studied in past years for both images and videos, and it has been demonstrated that SSIM is an effective measurement of perceptual global degradations [88]-[90]. However, it has been shown in [93] that SSIM is

very sensitive to contrast change, small translation and rotation between two images. This feature is an advantage when SSIM is used as quality metric in image processing. But in video compression, coding of a frame with slow motion or small contrast change with respect to its reference frame does not consume lot of bits with the help of efficient motion estimation/motion compensation and entropy coding. Therefore, SSIM cannot be directly used in video system for temporal complexity measure. An improved similarity index is needed so that it can be insensitive to small movement, contrast change but still sensitive to fast motion.

5.2.3.2 New Complexity Analyzer

In view of the aforementioned desired feature for temporal complexity measure, a new similarity index (*SMI*) is used to measure the similarity between two adjacent frames. Consider the k th and $(k + 1)$ th frame in a GOP. Let $z_k(p)$ and $z_{k+1}(p)$ be the luminance value at pixel position p in the two frames, respectively, where $p = ((x, y) | \{1 \leq x \leq L, 1 \leq y \leq W\})$. L and W represent the frame height and width in pixels. The similarity between frame k and $(k + 1)$ is calculated as [94]:

$$SMI(k) = \frac{2\mu_k\mu_{k+1} + c_1}{\mu_k^2 + \mu_{k+1}^2 + c_1} \cdot \frac{\sum_p \check{z}_k(p)\check{z}_{k+1}(p) + c_2}{\sum_p [\check{z}_k^2(p) + \check{z}_{k+1}^2(p)] + c_2} \quad (5.3)$$

where μ_k and μ_{k+1} are the average values of pixels in frame k and $(k + 1)$, c_1 and c_2 are the same as in Equation (5.2), $\check{z}_k(p)$ and $\check{z}_{k+1}(p)$ are the transformations from original frames $z_k(p)$ and $z_{k+1}(p)$ through

$$\check{z}_k(p) = z_k(p) - \Gamma(\beta) \cdot \mu_{z_k}, \quad (5.4)$$

$$\check{z}_{k+1}(p) = z_{k+1}(p) - \Gamma(\beta) \cdot \mu_{z_{k+1}}.$$

$\Gamma(\beta)$ in Equation (5.4) is a switch function and it is defined as:

$$\Gamma(\beta) = 1 - \beta^{(25(1-\beta(z_k, z_{k+1})))^{16}}, \quad (5.5)$$

$$\beta(z_k, z_{k+1}) = \frac{\sum_p z_k(p) z_{k+1}(p) + c}{\sqrt{\sum_p z_k^2(p) \sum_p z_{k+1}^2 + c}}. \quad (5.6)$$

where c is a small constant. $\beta(z_k, z_{k+1})$ is the cosine of the angle between the two lexicographic order vectors of frames k and $(k+1)$. The value of $\beta(\cdot)$ is in the range of $[0, 1]$, which serves as a preliminary detection of similarity between the two frames. It can be realized that when there exists big difference between the two frames, e.g. scene change, the value of $\beta(\cdot)$ approaches 0 that makes $\Gamma(\beta)$ approach 1. The SMI in Equation (5.3) becomes almost the same as SSIM accordingly, which can effectively detect the big difference. When the two frames are similar, e.g. small movement, the value of $\Gamma(\beta)$ would become 0. The true pixel values of the two frames are used to calculate the SMI in Equation (5.3) and it makes the SMI not sensitive to contrast change and small movement.

The value of SMI varies from 0 to 1 and a higher value means the two frames are more similar with each other or the complexity is low. Therefore, SMI is transformed to temporal complexity (TCX) and get the overall TCX of video i in GOP j with a total of K frames as follows:

$$TCX_{i,j} = \sum_{k=1}^{K-1} [1 - SMI(k)] \quad (5.7)$$

Spatial complexity measures the complexity in each frame. Kim *et al.* evaluated several methods in [96] on the complexity measure of still image, and found the gradient-based method is more reliable. Thus, the spatial complexity (SCX) of each frame is summed up to

define the spatial complexity $SCX_{i,j}$ of the GOP j in video program i as

$$SCX_{i,j} = \sum_{k=1}^K \left\{ \frac{1}{L \cdot W} \sum_{x=1}^{L-1} \sum_{y=1}^{W-1} [|z_k(x, y) - z_k(x+1, y)| + |z_k(x, y) - z_k(x, y+1)|] \right\}. \quad (5.8)$$

5.2.4 Joint Rate Allocation Scheme

5.2.4.1 Rate Allocation for Base Layer

In an SVC encoder, each video is encoded into a base layer and several FGS layers in each GOP. The base layer provides the minimum quality of the video, which is guaranteed to be transmitted when the bandwidth is low. Given the total transmission bitrate for base layers of all the videos, we need to distribute it among different videos before encoding. Wang *et al.* demonstrated in [85] that an efficient allocation of base layer bitrate is critical to achieve better overall performance of the whole system. They also implied in their experiment discussion that a more uniform base layer quality leads to a better inter-program fairness, i.e., all received programs have similar video qualities.

Due to the variation of scene complexity among videos, the base layer bitrate should be allocated to different videos in proportion to their complexity based on rate-distortion theory [86]. Both temporal and spatial complexity measures obtained in Section 5.2.3 are continuously sent to the joint rate allocator that dynamically distributes the total base layer coding bitrate. For the j th GOP in the i th video sequence, the allocated base layer bitrate is thus

$$BL_{i,j} = R_b \cdot \frac{TCX_{i,j} + SCX_{i,j}}{\sum_{i=1}^N (TCX_{i,j} + SCX_{i,j})}, \quad (5.9)$$

with R_b being the fixed total base layer coding bitrate and N being the total number of video programs in the system.

5.2.4.2 Rate Allocation for Video Adaptation

All the N bitstreams are multiplexed into a single channel with a channel bandwidth R_c . So the aggregate transmission bitrates of all the videos should be less than or equal to the channel capacity. Meanwhile, the channel should be fully utilized. So the rate constraint in every GOP is

$$(R_c - \epsilon) \leq \sum_{i=1}^N R_{i,j} \leq R_c \quad (5.10)$$

where $R_{i,j}$ is the transmission bitrate allocated to the j th GOP in the i th video, and ϵ is a predefined small value used as a threshold to adjust channel utilization level. ϵ is set to $0.002R_c$ in this work. In order to accurately control the transmission bitrate of every video bitstream, FGS is adopted in the system. It is possible to truncate an FGS bitstream at any byte position in the enhancement layers [95]. Given the total channel bandwidth, it is

FIGURE 5.2: Illustration of the problem in attaining both inter-program fairness and intra-program smoothness for broadcasting applications.

known that an equal distribution of the bandwidth results in a big quality difference across videos. Although Wang *et al.* [85] proposed a scheme to minimize the quality difference

between videos in one GOP, the smoothness between adjacent GOPs is not considered in [85]. As illustrated in Figure 5.2, the problem is a 2-D optimization problem where both inter-program fairness and intra-program smoothness should be considered. However, due to the non-stationary content of each video and the rate constraint in equation (5.10), at some time instants, it may be difficult to simultaneously obtain a smooth quality transition between GOPs in each video and a near-constant quality across all videos in each GOP. Therefore, the proposed scheme aims to dynamically allocate the bitrate to each video such that a good balance between intra-program smoothness and inter-program fairness can be achieved. The problem is formulated as

$$\begin{aligned} \min_{R_{i,j}^*} \{ & \omega \sum_{i=1}^N (D_{i,j} - \overline{D}_j)^2 + (1 - \omega) \sum_{i=1}^N (D_{i,j} - D_{i,j-1})^2 \} \\ \text{s.t. } & (R_c - \epsilon) \leq \sum_{i=1}^N R_{i,j}^* \leq R_c \end{aligned} \quad (5.11)$$

where $\omega \in [0, 1]$ is the smoothness factor used to adjust the importance of uniform quality among videos and smooth quality within one video. For the first GOP in the video, since no previous GOP is available, ω is set to 1 and only the inter-program fairness is considered. N represents the total number of video bitstreams multiplexed in the channel. $D_{i,j}$ and $D_{i,j-1}$ denote the distortion of the reconstructed video i in GOP j and $(j - 1)$, respectively. The average mean square error (MSE) is used to measure the distortion. $\overline{D}_j$ is computed as

$$\overline{D}_j = \frac{1}{N} \sum_{i=1}^N MSE_{i,j}. \quad (5.12)$$

The rate-distortion (R-D) function of SVC FGS layers is investigated and suggested to be a piecewise-linear model in [97]. Experiments were conducted on many sequences to verify the piecewise-linear R-D relationship in the FGS layers. Figure 5.3 illustrates the results obtained

FIGURE 5.3: R-D curve of different FGS layers.

from a random GOP with 16 frames of the sequence "Foreman". The base layer of the sequence is coded at QP=38 and the Hierarchical-B structure is used. It can be observed that the R-D curve of each FGS layer appears to be linear. Thus, knowing both end values of an FGS layer in an R-D curve, any in-between truncation points can be estimated using linear interpolation with the piecewise-linear model. In the proposed system, the coding statistics, e.g. the bitrate and distortion of the reconstructed video of base layer and all the FGS layers can be calculated as by-product during the encoding process. Let R_U and R_L be the upper and lower bound bitrates of a FGS layer, and their corresponding video distortion are D_L and D_U . For any

in-between truncation point, we have

$$\frac{D_U - D^*}{R^* - R_L} = \frac{D^* - D_L}{R_U - R^*} \quad (5.13)$$

where D^* is the video distortion of the truncation point and R^* represents the truncation bitrate. By manipulating equation (5.13) algebraically, we can get

$$R^* = \alpha \cdot R_U + (1 - \alpha) \cdot R_L \quad (5.14)$$

and

$$\alpha = \frac{D_U - D^*}{D_U - D_L}. \quad (5.15)$$

Based on the linear property of the bitrate and distortion model, a one-to-one mapping between the video bitrate and the video distortion can be obtained. Therefore, the searching of the optimal rate allocation $R_{i,j}^*$ in Equation (5.11) is equivalent to the searching of the corresponding video distortion $D_{i,j}^*$ of each program.

Although a brute-force search of $D_{i,j}^*$ can be used to solve the problem, it is not practical in reality because of the intensive computation. Hence, a simplified approach is used to find a sub-optimal solution to this problem. The details of the method are introduced in the following.

Step 1) Collect coding statistics and initialization

As explained in Section 5.2.2, the coding statistics such as bitrate and the corresponding video distortion of every SVC layer can be obtained as by-product during encoding process. These statistics are sent to Joint Rate Allocator as inputs of the bit allocation scheme. Let $R_l^{i,j}$ and $D_l^{i,j}$ denote the bitrate and distortion of the reconstructed

video after receiving l th layer and all the lower layers of $GOP_{i,j}$. $l = (0, 1, \dots, F)$ indexes over the SVC base layer and FGS layers. First, an initial search range $[a, b]$ is defined. It is a superset that includes all the distortion values from every encoder in the system. The values of a and b are thus determined as

$$a = \min(D_F^{i,j}), \quad i = 1, 2, 3, \dots, N \quad (5.16)$$

$$b = \max(D_0^{i,j}), \quad i = 1, 2, 3, \dots, N. \quad (5.17)$$

Step 2) Carry out golden-section search

An iterative golden-section search is carried out in this step. The main idea is to check the golden-section point c , i.e. distortion point positioned at $0.618(b - a)$ from the search interval endpoint in every iteration. All videos use current golden-section point c in each iteration as target distortion for bitstream truncation. If the sum of corresponding truncation bitrates could not meet the rate constraint in equation (5.10), the golden-section point c replaces one of the endpoints to form a new search interval. In detail, if $\sum_{i=1}^N R_c^{i,j} > R_c$, the new search interval in next iteration becomes $[c, b]$; If $\sum_{i=1}^N R_c^{i,j} < (R_c - \epsilon)$, the new search interval is $[a, c]$. This process repeats until a converged point D_{GS}^j is obtained. The corresponding truncation bitrates fulfills the constraint $(R_c - \epsilon) \leq \sum_{i=1}^N R_{GS}^{i,j} \leq R_c$, and for video i , since $D_{i,j} = D_{GS}^j$, the distortion variance across all the videos in GOP j is minimized.

Step 3) Carry out refinement search

After the converged golden-section point is obtained, refinement search is carried out in this step with a new search interval and a set of new search points. The number of search points is predefined. We firstly identify the video program whose distortion in

previous GOP has the biggest difference with D_{GS}^j , i.e.

$$I = \max_i |D_{GS}^j - D_{i,j-1}|, \quad i = 1, 2, 3, \dots, N. \quad (5.18)$$

The length of search interval is set to two times of the difference between D_{GS}^j and $D_{I,j-1}$.

If $D_{GS}^j > D_{I,j-1}$, the search range is $[A, B]$ with

$$A = D_{I,j-1} \quad (5.19)$$

$$B = 2 \times D_{GS}^j - D_{I,j-1}. \quad (5.20)$$

If $D_{GS}^j < D_{I,j-1}$, the new search interval would be $[B, A]$. The number of search points is predefined to 20 in this implementation and they are evenly distributed in the search interval. Every video takes a search point from the interval and all the combinations of search points are checked.

Step 4) Calculate the bitrate allocation

Every search point represents the target distortion of the bitstream after truncation. It is needed to locate the target distortion $D_{target}^{i,j}$ in one of the FGS layers of each video to find the corresponding bitrate $R_{i,j}$ as follows:

- If $D_{target}^{i,j} < D_F^{i,j}$, $R_{i,j} = R_F^{i,j}$.

- If $D_{target}^{i,j} \geq D_F^{i,j}$, $R_{i,j} = R_0^{i,j}$.

- If $D_{li,j} \leq D_{target}^{i,j} \leq D_{l-1}^{i,j}$, the target distortion is located in the l th FGS layer. The R-D curve of l th FGS layer is used to map the $D_{target}^{i,j}$ to the allocated bitrate $R_{i,j}$ as in equations (5.14) and (5.15). The sum of allocated bitrates is computed and only those

combinations fulfill the rate constraint in equation (5.10) are recorded. The combination which minimizes the cost function in (5.11) is the final bitrate allocation $R_{i,j}^*$.

5.2.5 Experimental Results

5.2.5.1 Experimental Setup

To evaluate the performance of the proposed system, a number of experiments were carried out on H.264/SVC reference software JSVM 9.8 [99]. Two sets of video sequences with different characteristics are used as video sources in the tests. All the sequences are International Standard testing sequences. Set 1 contains four video sequences at CIF (352×288 pixels) resolution. The sequences are "Foreman", "Football", "Mother & Daughter (M&D)", and "Harbour", and they are coded using Hierarchical-B structure with a GOP size of 16, frame rate of 30 fps. Set 2 consists of four HD (1280×720 pixels) video sequences, which are "Crew", "Bigship", "Raven", and "Shuttle_Start", with a framerate of 60 fps. They are coded with IPP...P structure and the GOP size is set to 30. A screen shot of each sequence is shown in Figure 5.4 and 5.5. Each sequence is also characterized in terms of the amount of motion, camera motion, texture and whether there is scene change as shown in Table 5.1.

FIGURE 5.4: Screenshot of CIF (352×288) sequences: "Foreman", "Football", "M&D" and "Harbor".

Section 5.2.5.2 evaluates the performance of base layer bit allocation, which demonstrates the accuracy of the proposed complexity measure. Complexity measure with better accuracy

FIGURE 5.5: Screenshot of HD (1280×720) sequences: "Crew", "Bigship", "Raven" and "Shuttle_start".

TABLE 5.1: Summary of different characteristics on test sequences

Sequence		Motion	Camera Motion	Texture	Scene Change
CIF	Foreman	L	M	L	Y
	Football	H	H	M	N
	M&D	L	N	L	N
	Harbour	M	N	H	N
HD	Crew	H	L	M	N
	Bigship	M	H	H	Y
	Raven	L	L	M	N
	Shuttle_start	L	L	L	N
*L = Low; M = Moderate; H = High; Y = Yes; N = No					

leads to more uniform base layer qualities. Section 5.2.5.3 investigates the performance of the proposed joint rate allocation scheme in terms of inter-program fairness and intra-program smoothness. Therefore, in all the experiments, the concern is not the PSNR in a particular GOP of a single video. Instead, we are looking at whether all the videos can achieve comparable qualities in each GOP and whether a stable quality can be achieved in each video. The performance comparisons are carried out between the proposed scheme and the scheme in [85], which has proven to be more efficient than many existing methods, such as the equal bitrate allocation scheme and the scheme in [84].

5.2.5.2 Results on Base Layer Bit Allocation

Base layer is the minimal guaranteed bandwidth of the channel. It is set to 1MB/s for Set 1 CIF sequences and 3MB/s for Set 2 HD sequences in the experiment. The total base layer

bandwidth is distributed to different videos for their base layer coding according to the relative video complexity.

This section evaluates the performance of base layer bit allocation by comparing the quality variance among videos in each GOP at their base layers. Variance is a measure of how far a set of numbers is spread out statistically. Thus a smaller variance represents more uniform base layer qualities. The quality variance for GOP j is calculated as

$$Variance = \frac{1}{N} \sum_{i=1}^N (PSNR_{i,j} - \overline{PSNR_j})^2 \quad (5.21)$$

where $N(= 4)$ is the total number of videos and $\overline{PSNR_j}$ is the averaged PSNR of N videos in GOP j . For each GOP, the PSNR of each video sequence at the base layer and the variance of PSNRs are displayed in Table 5.2 for CIF sequences, and Table 5.3 for HD sequences. The proposed scheme is compared with the scheme in [85], which measures the video complexity by combining I-frame complexity and P-frame motion activities. It can be seen that the variance of video qualities at base layer is reduced in every GOP using the proposed method. This is because as compared with [85], the proposed method allocates less bandwidth to the low complexity videos, such as "Foreman", "M&D" at CIF resolution, and "Shuttle_Start" at HD resolution. This results in more leftover base layer bandwidth that can be allocated to the high complexity videos. Consequently a more even base layer quality across all videos is attained in each GOP. If the channel bitrate drops and only the minimum base layer bandwidth is available in some extreme situations, quality differences among base layers directly indicates the inter-program fairness, and thus a better inter-program fairness can be achieved by the proposed method. On the other hand, since the rate is allocated according to relative complexity, the results also demonstrate that the proposed complexity measure is more accurate.

TABLE 5.2: Average PSNR & variance of PSNR for base Layers of different sequences at CIF resolution (Base Layer Bandwidth = 1MB/s)

GOP index	Average PSNR								Variance	
	Scheme in [85]				Proposed Scheme				Scheme in [85]	Proposed Scheme
	Foreman	Football	M&D	Harbour	Foreman	Football	M&D	Harbour		
1	32.22	27.07	34.08	28.87	30.49	27.93	34.36	28.49	7.54	6.36
2	32.34	27.29	34.65	28.74	30.85	28.48	34.65	27.88	8.44	7.06
3	32.49	28.21	35.04	27.84	30.40	29.10	35.47	26.88	9.07	9.95
4	32.47	28.69	34.67	27.41	31.14	29.31	34.18	27.41	8.44	6.22
5	32.28	30.30	33.36	27.97	30.39	31.35	34.33	27.48	4.22	5.99
6	33.99	35.19	35.20	28.35	32.66	34.53	35.20	29.60	8.03	4.72
7	34.00	34.50	35.27	28.43	33.02	33.65	35.27	29.20	7.32	4.96
8	33.82	33.24	35.71	27.74	32.34	33.78	34.86	27.74	8.81	7.38
9	32.97	30.98	34.61	27.34	30.72	31.85	34.06	27.34	7.35	5.88
10	33.26	30.56	33.80	26.88	30.13	31.60	32.75	27.68	7.51	3.58
11	34.02	29.25	34.53	26.91	30.29	30.10	33.52	27.47	10.31	4.60
12	35.68	29.89	34.90	26.57	32.99	30.49	34.90	27.74	13.91	7.23
13	34.89	29.52	34.78	27.20	31.41	30.54	34.41	27.81	11.17	5.55
14	34.01	26.71	35.27	26.76	31.44	27.65	34.26	27.64	15.82	7.77
15	33.38	26.86	35.08	26.72	32.06	27.63	34.51	27.06	14.22	9.61
16	32.56	27.46	35.74	26.92	31.39	28.08	34.60	27.12	13.39	8.68
17	31.46	28.33	35.65	27.51	30.39	28.73	33.72	28.02	10.23	4.84
18	31.67	28.30	36.65	27.62	30.73	28.77	35.54	27.53	12.76	9.29

5.2.5.3 Results on System Bit Allocation

In this section, the proposed algorithm is simulated with CIF sequences under the channel bandwidth of 2MB/s and with HD sequences under the channel bandwidth of 10MB/s. The bitrate for base layers is set as in previous experiment. Once the base layer coding bitrate for each video program is determined, each encoder compresses the video into a base layer and three FGS layers. The joint rate allocator dynamically allocates the truncation bitrate to each video and the bitstream extractor truncates accordingly. The proposed bitrate allocation scheme aims to minimize the cost function in (5.11) and balance between inter-program fairness and intra-program smoothness. ω is the parameter to adjust the importance between the two. A higher value of ω will make the algorithm more prone to uniform quality among videos. Table 5.4 and Table 5.5 list the average PSNR of each video and the variance of all

TABLE 5.3: Average PSNR & variance of PSNR for base layers of different sequences at HD resolution (Base Layer Bandwidth = 3MB/s)

GOP index	Average PSNR								Variance	
	Scheme in [85]				Proposed Scheme				Scheme in [85]	Proposed Scheme
	Crew	Bigship	Raven	Shuttle_Start	Crew	Bigship	Raven	Shuttle_Start		
1	34.56	30.94	35.95	39.17	35.73	32.09	33.08	38.85	8.71	6.87
2	34.02	32.71	34.27	39.95	34.68	32.17	34.65	38.80	7.76	5.67
3	35.30	32.17	34.65	39.25	35.38	32.67	34.52	38.05	6.46	3.75
4	33.95	33.01	34.25	38.61	34.10	33.01	34.25	37.79	4.66	3.23
5	33.29	32.09	35.13	37.29	33.52	32.86	34.29	36.64	3.86	2.04
6	32.86	32.80	33.87	37.15	33.39	32.98	33.73	36.01	3.14	1.38
7	31.35	30.64	33.35	36.75	32.80	29.92	32.58	35.72	5.63	4.21
8	31.24	31.41	32.31	35.39	31.24	31.71	32.31	35.21	2.79	2.39
9	31.05	30.75	31.21	35.11	30.98	30.93	31.44	34.68	3.19	2.42
10	30.16	29.77	33.39	34.62	30.26	30.20	31.87	34.25	4.30	2.71
11	28.98	29.07	32.78	34.20	30.42	29.07	31.75	33.42	5.24	2.59
12	29.24	28.56	31.61	34.38	29.47	28.98	31.06	33.40	5.22	2.98
13	30.74	28.96	31.80	33.98	30.51	29.60	30.97	33.08	3.30	1.63
14	29.73	28.84	31.83	32.19	29.81	29.27	31.60	31.41	1.97	1.00
15	30.16	29.57	31.46	33.56	30.42	29.96	30.40	32.81	2.35	1.26
16	30.27	29.80	33.19	34.56	30.45	30.18	30.52	33.68	3.95	2.05
17	30.87	29.93	34.12	34.77	30.80	30.58	31.61	33.86	4.26	1.69
18	30.58	30.17	33.47	35.40	30.82	30.80	30.31	34.39	4.61	2.67
19	29.82	30.25	32.00	34.84	30.70	30.05	31.63	33.14	3.89	1.34
20	31.89	31.20	33.74	36.50	32.24	31.72	32.19	31.38	4.22	0.13

the PSNRs in each GOP. The summation of quality fluctuations at GOP boundaries for each video is also shown in the tables for the two methods. The sum of fluctuation is calculated as

$$Sum\ Of\ Fluctuations = \sum_{j=2}^{N_{GOP}} |PSNR_{i,j} - PSNR_{i,j-1}| \quad (5.22)$$

where N_{GOP} is the total number of GOPs tested in the experiments.

The variance of PSNRs in each GOP reflects the performance of inter-program fairness. A smaller variance means a more uniform quality level and a better inter-program fairness is obtained. The sum of fluctuation is to show the quality variations in each video. A smaller fluctuation represents a better intra-program smoothness. The scheme in [85] only aims to

TABLE 5.4: Average PSNR , variance of PSNR in each GOP Of different sequences and sum Of PSNR fluctuations in each sequence (CIF resolution, Channel Bandwidth = 2MB/s, GOP = 16, Framerate = 30 frame/s, Hierarchical-B structure)

GOP index	Average PSNR								Variance (Inter-prog. fairness)	
	Foreman		Football		M&D		Harbour			
	Scheme in [85]	Proposed Scheme	Scheme in [85]	Proposed Scheme	Scheme in [85]	Proposed Scheme	Scheme in [85]	Proposed Scheme	Scheme in [85]	Proposed Scheme
1	32.22	30.91	30.66	31.07	34.08	34.36	30.84	31.11	1.88	2.09
2	32.34	30.90	30.62	31.14	34.64	34.64	30.75	31.03	2.63	2.45
3	32.49	31.03	30.93	31.48	35.47	35.47	30.81	31.08	3.53	3.45
4	32.47	31.04	30.98	31.32	34.67	34.67	30.79	31.11	2.41	2.33
5	32.38	31.10	31.94	31.85	33.36	34.33	31.71	32.01	0.40	1.46
6	33.99	32.81	35.19	34.97	35.20	35.20	32.94	33.21	0.88	1.10
7	34.00	32.82	34.50	34.83	35.27	35.27	32.80	33.00	0.80	1.17
8	34.05	32.86	33.24	33.78	35.71	34.86	33.03	33.03	1.11	0.62
9	32.97	32.41	31.80	32.36	34.61	34.14	31.81	31.81	1.33	0.76
10	33.26	32.41	31.76	32.21	33.80	33.75	31.76	31.73	0.82	0.56
11	34.02	32.29	31.31	31.90	34.53	33.99	31.10	31.32	2.39	0.99
12	35.68	33.19	31.71	32.00	34.90	34.90	31.61	31.86	3.37	1.48
13	34.89	32.88	31.51	31.84	34.78	34.45	31.25	31.34	2.99	1.42
14	34.01	32.24	29.96	30.78	35.27	34.52	29.88	30.33	5.70	2.67
15	33.38	32.06	30.07	30.49	35.08	34.51	29.92	30.29	4.85	2.84
16	32.56	32.31	30.51	30.70	35.74	34.60	30.25	30.07	4.82	3.06
17	31.46	32.29	31.18	31.23	35.65	34.60	30.97	30.75	3.75	2.21
18	31.67	32.08	31.22	31.21	36.65	35.54	30.98	30.76	5.43	3.52
Sum of Fluctuation (Intra-prog. smoothness)	10.51	5.06	11.91	9.62	11.37	7.40	7.96	7.21		

minimize the quality variance of all the videos in each GOP but it does not put any control on the quality variation in a particular video. Therefore, it can be seen that the proposed scheme achieves smaller quality fluctuation in every video sequence, whereas the variance of PSNRs in every GOP is still comparable as compared with the scheme in [85]. For many GOPs of CIF sequences as shown in Table 5.4, the proposed scheme even attains smaller variance than [85]. This is because the scheme in [85] inefficiently encodes the base layers of low complexity videos "Foreman" and "M&D" at higher bit rates than necessary due to inappropriate base layer bit allocation. Although more channel bandwidth is then allocated to the higher complexity videos, a reconstruction from their truncated bitstreams (comprising the base layer and partial enhancement layers) still results in substantially lower qualities than that of reconstructing low complexity videos whose bitstreams contain only a base layer. It

FIGURE 5.6: Quality fluctuation at GOP boundaries in each video sequence at CIF resolution with the channel bandwidth = 2MB/s

again verifies that a better base layer bit allocation with a more accurate complexity measure is beneficial for obtaining a better overall performance. In addition to the comparison on Sum of Fluctuations, the detailed quality fluctuation at each GOP boundary is plot out for all the videos in Figure 5.6 and Figure 5.7 to evaluate the intra-program smoothness. It can be seen that the proposed scheme gets smaller quality fluctuation at GOP boundaries in general and bigger improvement can be observed in the low complexity videos.

It should be acknowledged that the proposed scheme improves the system bit allocation with slightly increased computational complexity. As elaborated in Section 5.2.4.2, the increase of computational complexity as compared with the scheme in [85] mainly comes from the refinement search step for getting the target distortion $D_{i,j}^*$ of each video program. However, the proposed scheme has same codec delay as the scheme in [85].

TABLE 5.5: Average PSNR , variance of PSNR in each gOP Of Different sequences and sum Of PSNR fluctuations in each sequence (HD resolution, Channel Bandwidth = 10MB/s, GOP = 30, framerate = 60 frame/s, IPP...P structure)

GOP index	Average PSNR								Variance	
	Crew		Bigship		Raven		Shuttle_Start		(Inter-prog. fairness)	
	Scheme in [85]	Proposed Scheme	Scheme in [85]	Proposed Scheme	Scheme in [85]	Proposed Scheme	Scheme in [85]	Proposed Scheme	Scheme in [85]	Proposed Scheme
1	35.93	35.73	35.66	36.39	35.95	36.16	39.17	38.85	2.08	1.48
2	36.26	36.01	36.31	36.53	36.29	36.42	39.95	38.80	2.52	1.20
3	36.72	36.26	36.80	36.83	36.58	36.69	39.25	38.35	1.23	0.63
4	36.30	36.04	36.27	36.49	36.25	36.48	38.61	37.79	1.02	0.43
5	36.02	35.87	35.96	36.24	35.99	36.33	37.29	37.21	0.32	0.24
6	35.81	35.75	35.72	36.15	35.74	35.95	37.15	37.11	0.37	0.27
7	34.19	34.53	34.21	34.66	34.30	34.75	36.75	36.84	1.19	0.90
8	34.36	34.45	34.35	34.32	34.19	34.59	35.39	36.34	0.23	0.67
9	33.59	33.68	33.59	33.58	33.50	34.00	35.11	36.13	0.45	1.09
10	34.00	33.43	33.93	33.36	33.89	33.93	34.62	35.74	0.09	0.93
11	32.52	32.37	32.54	32.40	32.78	33.73	34.20	35.35	0.48	1.49
12	31.67	31.54	31.85	31.55	31.81	33.05	34.38	35.26	1.28	2.31
13	32.44	31.78	32.45	32.10	32.67	33.29	33.98	35.18	0.41	1.77
14	31.93	31.74	31.93	31.66	31.89	33.21	32.19	35.02	0.01	1.87
15	32.40	31.82	32.36	32.02	32.33	33.13	33.56	34.96	0.27	1.56
16	32.95	32.27	32.86	32.41	33.19	33.34	34.56	34.74	0.47	0.97
17	33.12	32.40	33.12	32.66	34.12	33.54	34.77	34.86	0.49	0.92
18	33.24	32.57	33.42	32.85	33.47	33.52	35.40	34.81	0.77	0.75
19	32.94	32.65	32.95	33.08	32.79	33.63	34.84	34.35	0.71	0.40
20	34.19	33.14	34.29	34.35	34.10	34.87	36.50	35.10	1.01	0.57
Sum of Fluctuations (Intra-prog. smoothness)	11.12	6.92	11.49	9.40	12.69	6.33	14.35	5.49		

5.2.6 Summary

A joint rate allocation scheme for multiprogram video coding is proposed in this section. The scheme is developed on a platform where the scalable extension of H.264/AVC is applied to compress the video programs in order to avoid complicated re-encoding and transcoding. A new complexity measure that incorporates the characteristics of Human Visual System is introduced for base layer bitrate allocation. With benefit from a better base layer bit allocation, the system achieves a more uniform quality level for all multiplexed videos when channel bitrate is low. A new 2-D channel bitrate allocation proposed in the section is designed on the

FIGURE 5.7: Quality fluctuation at GOP boundaries in each video sequence at HD resolution with the channel bandwidth = 10MB/s

basis of ensuring a comparable quality level for all videos while minimizing the quality variations within each video. It also provides a way to adjust the balance between smoothness and fairness performance. The overall performance of the proposed system has been evaluated via simulations and compared with existing methods. Experimental results show that the proposed system obtains a notable decrease in the intra-program quality fluctuation, and still maintains a comparably good inter-program fairness. In applications of video broadcasting, the proposed scheme is very useful to provide a high degree of comfort by bringing minimum quality fluctuations to clients either when they switch from one TV program to another, or just stay in the same TV program.

5.3 Joint Rate Allocation of Stereoscopic 3D Videos in Next-Generation Broadcast Applications

Section 5.2 addressed the bit allocation of multiple video in current broadcast applications. With recent advances in display and camera technology, 3D video is foreseen to be the next natural step in the evolution of digital media and we are now seeing a global roll-out of services to support 3D video on consumer electronics platforms. Since 3D video is getting popular, the problem addressed in Section 5.2 should also be studied for next-generation broadcast applications, such as 3DTV, where multiple 3D videos are encoded in parallel and transmitted simultaneously over a bandwidth-limited broadcast channel. Due to the increased video data compared to conventional single-view video, efficient presentation format and coding technology are the first important issues to consider. The bit allocation problem for 3D videos is also more complicated as it includes both bit allocation between videos as well as between views.

5.3.1 Background and Related Work

In this section, a short overview of available options for 3D video representation and coding is given, followed by an introduction of asymmetric stereo video coding method that used in the design of the proposed rate allocation scheme.

5.3.1.1 3D Video Representation and Coding

Current 3D video can be generally classified into stereoscopic and multiview. Stereoscopic video presents two offset views separately to left and right eyes to create the illusion of depth. Multiview video includes more views as compared with stereo format. It is used to provide

free-view functionality that enables viewers to perceive 3D video from different viewing angles. Three representation formats are commonly used for storage and transmission of 3D videos.

(1) Full Resolution Representation Format

Full-resolution format refers to store video frames without any change in spatial and temporal resolution. For stereo video, the frames of left and right views are stored sequentially to support full HD resolution. However, this format doubles the raw data rate for stereo video, and there will be an N -fold increase in the data rate for multiview video with N views.

(2) Frame-Compatible Representation Format

Frame-compatible is another common format to enable stereoscopic service through the existing hardware and delivery channel. It has been used to provide 3DTV service over existing digital TV broadcast infrastructures [102]. Frame-compatible format employs subsampling to keep the frame size and data rate the same as that of 2D videos. There are various options for subsampling and common patterns include side-by-side, top-and-bottom, line interleaved and checkerboard. Among these formats, side-by-side and top-and-bottom are selected as mandatory format for broadcast by the latest HDMI specification 1.4a [103]. However, an obvious drawback of frame-compatible format is that the spatial or temporal resolution would be lost. Figure 5.8 shows an example of side-by-side and top-and-bottom frame-compatible format as compared with full-resolution.

(3) Depth-Based Representation Format

Depth-based format is another important representation for 3D video, where views and their associated depth maps are transmitted to render/synthesize intermediate virtual views through depth-based image rendering (DBIR) techniques. The depth information on depth

FIGURE 5.8: (Top) Full-resolution left and right views, and (Bottom) frame-compatible representation of stereoscopic videos.

map represents the distance of each video pixel from the camera. The depth values are scaled to 8 bit so the depth map can be considered as a gray-scale video. This format can support both stereo and multiview displays in principle. ISO/IEC 23002-3 [104] has specified a container format for auxiliary video and supplemental information in view-and-depth data. This format was later extended to multiview video plus depth, where N views and N depth maps are transmitted to decoder to generate M views with $M \geq N$. View synthesis requires extrapolation of image content to some extent. On the synthesized virtual view, parts of the 3D scene may become visible that are occluded behind foreground objects in the available original video. If the virtual viewpoint is close to the original camera position, masking of uncovered image regions works well with limited artifacts. However, with increasing distance of the virtual viewpoint also the extrapolation artifacts increase. The challenging problems of depth-based format are thus the generation of accurate depth maps and the view synthesis at occluded areas.

Different coding formats have been developed for 3D videos but there is no standardized format for storage and transmission system. The coding formats for 3D videos should depend on the representation format and transport options. Different coding formats that are available for aforementioned three representation formats are presented in the rest of this subsection.

(1) Coding format for stereo video

A. Simulcast coding

Stereo videos come with two offset views. The most straightforward way to compress the two views is simulcast coding, i.e., to code each view independently using a state-of-the-art video codec. This method keeps processing delay to a minimum and it provides backward compatibility spontaneously when only one view is decoded. However, the main drawback of simulcast is that coding efficiency is low because the redundancy between two views is not exploited.

B. Frame-compatible coding with SEI message

As introduced earlier, frame-compatible format employs subsampling to keep the frame size and data rate the same as that of 2D videos. Thus frame-compatible videos can work on existing infrastructures and can also coded by existing codecs. H.264/AVC [6] specified a new Supplemental Enhancement Information (SEI) message that enables the signaling of the frame-compatible video used. This SEI message contains information of video format, sampling relationship and view order of the two views. Decoder can immediately recognize the format and when frame-compatible is detected, necessary processing associated with the frame-compatible format will be carried out.

C. Frame-compatible on SVC

FIGURE 5.9: Frame-compatible coding using SVC with spatial scalability.

Although the frame-compatible format can be deployed to existing platforms, it still suffers from reduced resolution and thus reduced quality perception. Recently several work have been proposed to extend frame-compatible signal to full resolution [105]. An obvious method to achieve full resolution is to make use of the Scalable Video Coding (SVC) extension of H.264/AVC. For example, one may use the spatial scalability in SVC to code a lower resolution frame-compatible signal as base layer while full resolution frame-compatible signal is coded as enhancement layer. This method is shown in Figure 5.9 with the side-by-side subsampling as an example. V_0 and V_1 present two views from the stereo videos, arrows indicate the prediction relationship between frames, and t denotes the frame index. Another method that also makes use of SVC utilizes both spatial and temporal scalability is shown in Figure 5.10. Only half of the frame-compatible frame is upsampled in first enhancement layer to provide full resolution of one view. The full resolution of second view can be encoded as a temporal enhancement. However, both methods have the drawback that they may not be practical for more complicated frame-compatible formats, such as frame-compatible using checkerboard subsampling format.

(2) Coding format for multiview video

Since there are more than two views in multiview video format, it is very important to employ a high efficient video codec to properly remove the redundance between views. The

FIGURE 5.10: Frame-compatible coding using SVC with a combined spatial and temporal scalability.

H.264.AVC has published an amendment to support Multiview Video Coding (MVC) [6]. In MVC, decoded frames from each view are made available for prediction of other views through flexible reference list management. Every frame can thus be predicted from its inter-view reference or temporal reference. Another important feature of MVC is that, the base view is compatible with H.264/AVC. By extracting and decoding only the base view of the stream, backward compatibility can also be achieved.

(3) Coding format for depth-based video

With the feature of exploiting inter-view redundancy, MVC is usually adopted for coding of depth-base video. Depth map associated with each view can be considered as a grey-scale image that mainly comprises large homogeneous areas inside objects and sharp transitions around objects at different depths. Since blurring artifacts appear in the reconstructed video at high compression ratio, the sharp edges on depth maps usually degrade a lot too in this case. Since depth map is used for intermediate view rendering, the coding errors in reconstructed depth maps would cause annoying artifacts in rendered views. Therefore, the different characteristics of depth map should be taken into consideration and the sharp edges in depth

map should be preserved better than video data. Schemes have been developed to encode depth-based video by applying rate-distortion optimization principles on synthesized views too [106]. Several edge-aware synthesis algorithms can also be applied as post-processing steps to reduce the errors caused by depth coding with MVC [107][108].

5.3.1.2 Asymmetric Stereo Video Coding and Quality assessment

Although there are several representation formats available for 3D video, stereoscopic seems to be the most mature technology at this stage [100]. This work thus focuses on stereo video. Simulcast coding of stereo video keeps computation delay to a minimum but suffers from low coding efficiency due to the redundancy exists in two coded views. Based on the characteristics of human visual system (HVS), it has been shown that human eyes can tolerate and conceal the missing high-frequency information in one of the views [110]- [114]. Therefore, to reduce the overall transmission rate of stereo videos, many algorithms on asymmetric stereo coding have been proposed to reduce the fidelity in one of the views. It is possible to achieve asymmetry through spatial [115], quality [109][110] or temporal [116] scaling in one of the views. Both spatial asymmetry and quality asymmetry can be managed to achieve unnoticeable degradation in perceptual 3D quality [110][101], but it is also found that the temporal asymmetry may bring in unacceptable perceptual quality for high motion contents [116].

Stelmach *et al.* [113] and Tam *et al.* have been among the first to study quality assessment of asymmetric coding. Questions regarding which scaling method should be used to achieve asymmetric and what level of scaling can be applied before observers can notice degradation are the key concerns. The effect on spatial and quality asymmetry has been investigated in [112][113]. Their subjective test results favored spatial scaling over quality scaling because they found blockiness artifact caused by quality scaling is more disturbing than blur artifact

caused by low-pass filtering used in spatial scaling. More recent studies argued that spatial scaling makes the perceived quality coarsely changed and thus suggested using quality scaling to enable finer tuning in the level of quality asymmetry [118][119]. Subjective test results shown in [118] also concluded that quality asymmetry is preferred over spatial scaling on full resolution 3D display, which appears to be contradicted with [112][113]. The cause of the contradiction has not been determined but different codecs and display units used in the subjective tests might be one of the causes[101]. Among different approaches of asymmetric coding, quality asymmetry seems to have the most straightforward implementation and the results still look promising. It was reported in [118][119] that unnoticeable perceptual quality degradation can be obtained if the quality (PSNR) of the low quality view is kept higher than a "just-noticeable" threshold. And this threshold is then proved to be "33dB for the polarized projection display and 31.5dB for the parallax barrier display" by [101]. At the same time, [101] also concluded that at low bit rate where both views cannot get PSNR higher than the just-noticeable threshold, symmetric coding should be used.

Currently subjective viewing test is the common approach to assess the quality of stereo video, because there is no quantitative metric proposed in literature that has been widely accepted for 3D quality evaluation. Although PSNR is a quality metric that is frequently used and widely adopted for conventional 2D video, simple averaging of the two PSNR values from two views without considering the characteristics of HVS can cause misleading results. It is known that HVS can tolerate one of the views at unequal quality, it is thus more difficult to correlate the PSNR value with perceived quality. In [120], the proposed metric makes use of rate distortion model to estimate visual quality with PSNR and level of jerkiness. The influences of asymmetry in all the possible scaling dimensions, i.e., spatial, temporal and quality dimensions, have been analyzed in this metric. [121] proposed a metric using

spatial frequency dominance model based on the idea that the degraded view should not have high spatial frequency components. The metric is calculated using mean absolute differences (MAD) and spatial frequencies imparities between degraded and original view. Although the proposed metric in [121] shows high correlation with subjective evaluation results MOS, the values obtained by the metric are not as intuitive as PSNR. A simple yet effective metric that correlate with PSNR was proposed in [118]. the metric was designed based on the conclusion from subjective test that when PSNR of both views are close to each other, the higher PSNR reflects the perceived visual quality better. The metric is shown below,

$$Q(\alpha, \beta) = \begin{cases} PSNR_{HIGH} & \text{if } \frac{PSNR_{LOW}}{PSNR_{HIGH}} > \alpha \\ \beta \cdot PSNR_{HIGH} + (1 - \beta) \cdot PSNR_{LOW} & \text{otherwise} \end{cases} . \quad (5.23)$$

where Q represents the perceived quality. α is a threshold to evaluate whether the PNSR from two views are close to each other, and it was proposed to be 0.85. β is a weighting parameter that depends on the display technology and spatial resolution of the degraded view.

5.3.2 System Overview

The basic structure of joint coding system for multiple stereo videos is shown in Figure 5.11. The main components in the system comprise several preprocessors, scalable stereo coding (SSC) encoders, bitstream operators, joint rate allocator and a multiplxer. The structure is very similar to the one as shown in Figure 5.1 in Section 5.2.2 but functionalities of some components are different in this system.

The input of this system is multiple stereo videos with two views in each video. First, each stereo video goes through a preprocessor, which is the same as the complexity analyzer in Figure 5.1, for analysis of the video content on a GOP basis. The picture statistics are derived

to characterize the overall complexity of the two views in each GOP. The joint rate allocator collects all the statistics from preprocessors and uses this data for base layer bitrate allocation among videos based on their relative complexity. The base layer bit rate represents the minimum guaranteed bandwidth of the channel. After receiving the input video as well as the information of allocated base layer bitrate, each SSC encoder can encode the stereo video into a scalable bitstream including two scalable portions of the two views. Details on the proposed SSC encoder and the generated scalable bitstream will be given in subsection 5.3.3. Since the encoding is performed, the coding statistics can be obtained as byproduct from the encoding process. The joint rate allocator takes these coding statistics and calculates the optimal bitrate allocation among different videos as well as bitrate allocation between two views of each video in a GOP. The bitstream and information of bitrate allocation are then fed into bitstream operator which first extracts two substreams corresponding to the two views from the original bitstream based on the bitrate allocation. The bitstream operator reforms a new bitstream by putting the two substreams together. In the multiplexer, the reformed bitstream of different videos are multiplexed and transmitted over the channel. Please note that, at decoder side, the bitstream received should be preprocessed to create two separate bitstreams for the decoding of stereo views.

5.3.3 Scalable Stereoscopic Video Coding

In this subsection, a new coding scheme, namely scalable stereo coding, is proposed for stereo videos. The coding formats for 3D videos should depend on the representation format and transport options. Stereo video format is the focus in this work because stereoscopic video seems to be the most mature representation format at this stage [100]. The two views in stereo video are usually called reference view and auxiliary view, where the reference view is

FIGURE 5.11: Basic structure of joint coding system for multiple stereo videos

the conventional monoscopic 2D video used in existing infrastructure and the auxiliary view refers to the additional view included in stereo video. Considering the joint bit allocation problem to be studied, we think the coding format of stereo videos should fulfil following requirements:

- The bitstream corresponding to both views should be scalable. It is to avoid complicated re-coding or transcoding when performing rate control of the video. With a scalable bitstream, accurate rate control can be easily achieved through bitstream extraction.
- The coding scheme should try to remove redundancy between two views in order to improve the coding efficiency. This can be done by enabling inter-view prediction.
- Since bitstream could also be truncated for the reference view, the prediction structure between two views should be designed in a way such that the drift between encoder and decoder can be minimized.
- Backward compatibility should be provided.

Please note that since majority of current display systems use polarized full resolution projection, resolution change is not considered in the schemes and all the stereo videos used in this work are kept at their original resolutions. The coding structure of the proposed SSC is depicted in Figure 5.12. Arrows in the figure indicate the direction of prediction. Reference view is first coded as a base layer by SVC. Every frame in auxiliary view can be either predicted from its corresponding frame in reference view or its neighboring frame in temporal domain. The temporal prediction among frames in the same view can be the same as that in the state-of-the-art H.264/AVC where either "Hierarchical-B" or "IBP"/"IPP" structure can be used. The scalability of each view is provided by enabling progressive quality refinement, such as FGS that refines the residual to supports a continuous scalable bitrate range. Since bitstream truncation of the reference view can be carried out at any scalable portion, the auxiliary view is only predicted from the base layer of the reference view. Although it sacrifices some coding efficiency because the prediction doesn't use the highest quality reference, it completely avoids drifting.

FIGURE 5.12: Scalable stereo coding structure.

The structure of the scalable stereo bitstream with L quality refinement layers for both views is shown in Figure 5.13. Even though auxiliary view is coded similarly as an enhancement layer

FIGURE 5.13: Structure of scalable stereo bitstream with L quality refinement layers for both views.

in SVC (e.g. predicted from reference view), it is at the same spatial resolution as reference view and its quality is not necessarily to be better. We thus still mark it as "base layer" in Figure 5.13. If the overall bitrate allocated to j th GOP in the i th stereo video is $R_{i,j}$, and the bitrates allocated to two views are R_{ref} and R_{aux} , respectively (i.e. $R_{ref} + R_{aux} = R_{i,j}$). The bitstream operator as shown in Figure 5.11 first truncates the bitstream into two substreams that match the two allocated bitrates. So we can get

$$r_{ref}^T = R_{ref} \quad (5.24)$$

$$r_{aux}^T - r_{aux}^0 = R_{aux} \quad (5.25)$$

where r represents the position on bitstream in Figure 5.11. The two substreams are combined together by bitstream operator for transmission. At decoder side, the decoder parses the transmitted bitstream and separates it into two substreams, with sizes R'_{ref} and R'_{aux} , for the decoding of two views, respectively. Since auxiliary layer is predicted from reference view, base layer of reference view must be added to the auxiliary bitstream for decoding of auxiliary view. Thus the two substreams used for decoding are formed as

$$R'_{ref} = R_{ref} = r_{ref}^T \quad (5.26)$$

$$R'_{aux} = R_{aux} + r_{ref}^1 - r_{ref}^0. \quad (5.27)$$

5.3.4 Joint Rate Allocation for Multiple Stereo Videos

The joint bit allocation problem for stereo videos is more complicated as it includes both bit allocation among videos as well as between views in each video. In this subsection, the distribution of channel bandwidth among videos is first discussed in Subsection 5.3.4.1. It includes two steps. First step is to distribute the minimum guaranteed channel bandwidth among videos so that each encoder can use this information to encode the video into a bitstream including base layer and quality refinement layers. Second step determines how to distribute overall channel bandwidth so that the perceived quality of each video by decoding the truncated bitstreams are almost the same. Since there are two views for stereo video, bit allocation among two views is studied in each video to get higher perceived quality in Subsection 5.3.4.2.

5.3.4.1 Rate Allocation Among Videos

Due to the fact that both views must be available for decoding to create 3D perception, the minimum guaranteed channel bandwidth distributed to each video is used to code the base layers of reference view and auxiliary view. Based on the study in Subsection 5.2, we know that the minimum guaranteed bandwidth should be distributed in proportion to relative complexity of each video. Thus the complexity of each video is obtained by summing up the temporal complexity (TCX) and spatial complexity (SCX) of both views. The TCX and SCX of each view are obtained using the proposed complexity measure in Subsection 5.2. The information of TCX and SCX are continually sent to joint bit allocator which dynamically distributes the minimum guaranteed channel bandwidth R_b to j th GOP of the i th stereo video

as

$$BL_{i,j} = R_b \cdot \frac{TCX_{i,j} + SCX_{i,j}}{\sum_{i=1}^N (TCX_{i,j} + SCX_{i,j})}, \quad (5.28)$$

where N denotes the total number of stereo videos in the system. $TCX_{i,j}$ and $SCX_{i,j}$ are the sum of corresponding complexity from both views. Knowing the allocated $BL_{i,j}$, each encoder can start the encoding process to generate a bitstream as shown in Figure 5.13. The size of $(r_{ref}^1 - r_{ref}^0) + (r_{aux}^1 - r_{aux}^0)$ is equal to the allocated $BL_{i,j}$.

Next step is to distribute channel bandwidth among videos. All the N bitstreams are now available for truncation before they are multiplexed and transmitted through a single channel. So the aggregate transmission bitrate of all the videos should be less than or equal to channel bandwidth. Meanwhile the channel should be fully utilized. So we have the rate constraint for all the stereo videos in every GOP as

$$(R_c - \epsilon) \leq \sum_{i=1}^N (R_{ref}^{i,j} + R_{aux}^{i,j}) \leq R_c \quad (5.29)$$

where $R_{ref}^{i,j}$ and $R_{aux}^{i,j}$ are the bitrate allocated to reference view and auxiliary view, respective, for the j th GOP in i th video.

In broadcast applications, keeping almost uniform quality among multiple video programs is very important. It provides high level user experience when clients switch from one video program to another. Therefore, the objective here is to dynamically allocate channel bandwidth R_c to each video such that the variation of perceived quality among videos will be

minimized. The problem can be formulated as

$$\begin{aligned} \min_{R_{ref}^{i,j}, R_{aux}^{i,j}} Var &= \frac{1}{N} \sum_{i=1}^N (Q_{i,j} - \overline{Q_j})^2 \\ s.t. (R_c - \epsilon) &\leq \sum_{i=1}^N (R_{ref}^{i,j} + R_{aux}^{i,j}) \leq R_c \end{aligned} \quad (5.30)$$

with Var being the variance of perceived quality among different stereo videos, $Q_{i,j}$ being the perceived quality of i th video in j th GOP, and $\overline{Q_j} = \frac{1}{N} \sum_{i=1}^N Q_{i,j}$ being the mean of perceived $Q_{i,j}$ for N videos. ϵ is a predefined small value used as a threshold to adjust channel utilization level. ϵ is set to $0.002R_c$ in this work. Minimizing the variance of perceived video quality, in another words, is to obtain almost uniform quality in all the videos, i.e. inter-program fairness.

5.3.4.2 Adaptive Inter-view Rate Allocation

It can be seen from the cost function in (5.30) that, it is needed to move on to inter-view bitrate allocation, finding the set of $R_{ref}^{i,j}$ and $R_{aux}^{i,j}$ for each video such that the estimated quality in reference view and auxiliary view can achieve a perceived quality $Q_{i,j}$ that minimizes Var . Please note that the estimated quality of each view is used in the process because an accurate quality can only be obtained at decoder side after the reconstructed video is available. Since coding statistics (i.e., R-D relationship) have been sent to the joint rate allocator during encoding, the joint bit allocator can estimate the quality of each truncated bitstream. The R-D relationship has been verified to be piece-wise linear in Section 5.2 and the plot of verification results is also shown in Figure 5.3. Thus an one-to-one mapping can be obtained between bitstream truncation point and the estimated quality for each view.

In Equation (5.30), if the perceived quality of N videos can be the same, variance can be minimized (i.e., $Var = 0$). Therefore, to solve the optimization problem in (5.30), it is equivalent to search for an optimal perceived quality, denoted as Q_j^* , such that when all videos in the system are estimated to achieve this targeted perceived quality, the two views of each video are truncated with $R_{ref}^{*i,j}$ and $R_{aux}^{*i,j}$ ($i = 1, 2, \dots, N$) that fulfill rate constraint in Equation (5.29).

There are thus two problems to be solved here:

- (1) How to evaluate the perceived quality of stereo video.
- (2) How to allocate bit rate between two views so that the targeted perceived quality can be achieved with high coding efficiency.

For perceived quality measure, the simple yet effective quality metric proposed in [118] is used and it is shown in Equation (5.23). The quality metric in [118] adopts the higher PSNR as perceived quality if PSNRs of both views are close to each other (if the ratio > 0.85). But if the two PSNR are very different, the low quality view starts to degrade the perceived quality and the perceived quality is a weighted some of two PSNRs. As introduced earlier in this section, asymmetric coding with quality scaling on auxiliary view is a good choice to save bits. In addition, Saygili *et al.* concluded in [101] that unnoticeable perceptual quality degradation can be achieved if the quality of the auxiliary view is kept higher than a just-noticeable threshold; This threshold is "33dB for the polarized projection display and 31.5dB for the parallax barrier display"; If both views cannot get PSNR higher than the just-noticeable threshold at low bit rate, symmetric coding should be used. With these considerations, the bitrate allocated between views should be adaptive depending on the targeted perceived quality Q_j^T . The adaptive inter-view bitrate allocation scheme is summarized below:

- If Q_j^T is less than just-noticeable threshold, symmetric coding should be used to get better perceived quality. Equal inter-view rate allocation is used so that the truncated

bitstream of both views are estimated to get an PSNR equals to Q_j^T .

- If Q_j^T is higher than just-noticeable threshold, asymmetric coding should be used to get high perceived quality with high coding efficiency. Unequal inter-view rate allocation is designed:
 - If the ratio between just-noticeable threshold and Q_j^T is larger than threshold α , the rate is allocated unequally such that the estimated quality of reference view can be Q_j^T while the estimated quality of auxiliary view equal to just-noticeable threshold only.
 - If the ratio between just-noticeable threshold and Q_j^T is smaller or equal to threshold α , the rate is allocated unequally such that the estimated quality of reference view is Q_j^T but the estimated quality of auxiliary view must be increased to a minimum value whose ratio with Q_j^T is equal to threshold α .

In the process of finding an optimal perceived quality Q_j^* , many sets of $R_{ref}^{i,j}$ and $R_{aux}^{i,j}$ can be obtained through inter-view rate allocation scheme. Only the set fulfills the rate constraint in Equation (5.29) is the final $R_{ref}^{*i,j}$ and $R_{aux}^{*i,j}$.

5.3.4.3 Algorithm Flowchart

All the encoders can encode the videos after the allocation of minimum guaranteed channel bandwidth. The coding statistics, especially the R-D relationship in scalable quality refinement layers, are obtained during encoding as byproduct. To solve the optimization problem in (5.30), it is a process to find optimal Q_j^* and the bit allocation results of $R_{ref}^{*i,j}$ and $R_{aux}^{*i,j}$. Although exhaustive searching can be applied to solve the problem and get optimal result, it is not feasible due to the large computation. Thus a sub-optimal fast search algorithm,

golden-section search method [122], is used to solve this problem. The flowchart of the joint rate allocation of N videos at the j th GOP is depicted in Figure 5.14.

In the initialization stage, the initial search interval $[L_1, U_1]$ is a superset of scalable quality ranges of all the videos. Assuming there are F quality refinement layers for each view and 0th layer represents the base layer. The initial search interval is determined as

$$L_1 = \min[PSNR_{ref}^{i,j}(0), PSNR_{aux}^{i,j}(0)], \quad i = 1, 2, \dots, N \quad (5.31)$$

$$U_1 = \min[PSNR_{ref}^{i,j}(F), PSNR_{aux}^{i,j}(F)], \quad i = 1, 2, \dots, N \quad (5.32)$$

In golden-section search method, the golden-section (GS) point in the search interval of each round is set to be the targeted perceived quality Q_j^T for evaluation. In round k , the GS point g_k is positioning at $0.618 \times \text{search interval}$ from the lower bound, i.e.

$$g_k = L_k + (U_k - L_k) * 0.618 \quad (5.33)$$

The adaptive inter-view rate allocation steps are carried out as shown in Figure 5.14, where JNT represents the just-noticeable threshold, and α is the ratio used to determine whether the two estimated PSNRs of two views are close to each other. If Q_j^T is less than JNT, $R_{ref}^{i,j}$ and $R_{aux}^{i,j}$ are set to the values such that the corresponding estimated $PSNR_{ref}^{i,j}$ and $PSNR_{aux}^{i,j}$ can be equal, which is equivalent as symmetric stereo coding. If Q_j^T is larger than just-noticeable threshold, rate allocation is asymmetric and the ratio between estimated $PSNR_{aux}^{i,j}$ and $PSNR_{ref}^{i,j}$ must be larger than or equal to the threshold α . The total size of truncated bitstreams from N videos is then checked with rate constraint defined in Equation (5.29) in each round. If the constraint cannot be fulfilled, the search interval and GS points are refined

as shown in the flowchart, else the process of rate allocation is completed. In this work, 33dB is set to be JNT as in [101] and α is chosen to be 0.85 as in [118].

FIGURE 5.14: Flowchart for finding the optimal bit allocation by the golden section search method.

FIGURE 5.15: Screenshot of stereo videos: "Street", "Fountain", "Lake", "Rollerskate" (a) Left View (b) Right View.

5.3.5 Experimental Results

To assess the performance of the proposed system, a number of experiments were carried out with 4 stereo videos, namely "Street", "Fountain", "Lake" and "Rollerskate", in the system for joint rate allocation. The resolution of each view is 640×480 and screenshots of the four stereo videos are shown in Figure 5.15. The proposed algorithm was implemented on JSVM [99] but the software has been modified such that the auxiliary layer is only predicted from the base layer of the reference view in order to avoid drift. The bitstream extraction and decoding process of JSVM have also been modified. The decoder creates two separate reconstruction videos for the two views of each stereo video. The videos are coded on GOP basis with a GOP size of 16. The joint bit allocation is updated at GOP boundaries.

The minimum guaranteed channel bandwidth is set to 5MB/s in the experiments. The proposed algorithm was tested under two channel bandwidths of 20MB/s and 40MB/s, respectively. To evaluate the performance, both PSNRs of the two views were recorded and the perceived PSNR was also computed based on the quality metric proposed in [118]. However, the metric in [118] does not consider the JNT, which means the perceived quality metric

needs to be modified. From the proposed joint bit allocation described in previous subsection, it can be seen that the symmetric rate allocation is automatically adopted when JNT cannot be reached, the quality of two views should be almost the same in this case. Therefore, the perceived quality is proposed to be the average of two views in this case. The modified computation of perceived quality Q is shown below in Equation (5.34).

$$Q = \begin{cases} PSNR_{HIGH} & \text{if } PSNR_{LOW} > JNT \& \frac{PSNR_{LOW}}{PSNR_{HIGH}} \geq \alpha \\ \beta \cdot PSNR_{HIGH} + (1 - \beta) \cdot PSNR_{LOW} & \text{if } PSNR_{LOW} > JNT \& \frac{PSNR_{LOW}}{PSNR_{HIGH}} < \alpha \\ \frac{1}{2}(PSNR_{HIGH} + PSNR_{LOW}) & \text{if } PSNR_{LOW} < JNT \end{cases} \quad (5.34)$$

Performance comparison was carried out between the proposed scheme and another two commonly used frameworks, the side-by-side frame-compatible representation format [105] used for 3DTV broadcast on DTV platform [102] and the conventional statistical multiplexing framework with full-resolution representation format, both coded by the state-of-the-art H.264/AVC. Channel bandwidth is equally allocated to frame-compatible videos in existing system, thus the same way of equal distribution is used in the experiment. For statistical multiplexing framework, the channel bandwidth is allocated in proportion to relative complexity of each video and the complexity measure proposed in Section 5.2.3 is used here, which is the same complexity measure used to allocate minimum guaranteed bandwidth in the proposed scheme. The objective of joint rate allocation scheme is to achieve almost uniform perceived quality among all the videos in the system so as to improve the user experience in broadcast applications. Thus the variance of all the perceived quality in each GOP is also compared. The variance is computed as

$$Variance = \frac{1}{N} \sum_{i=1}^N (Q_{i,j} - \overline{Q_j})^2. \quad (5.35)$$

TABLE 5.6: PSNR in individual view, perceived quality, and variance of perceived quality for 4 stereo videos achieved by proposed scheme (Channel bandwidth = 20MB/S)

GOP Index	Street		Fountain		Lake		Rollerskate		P-Quality Variance
	V-PSNR	P-Quality	V-PSNR	P-Quality	V-PSNR	P-Quality	V-PSNR	P-Quality	
1	32.42	32.36	32.55	32.43	32.32	32.42	32.41	32.37	0.001
	32.30		32.30		32.51		32.33		
2	32.24	32.24	32.43	32.40	32.87	32.25	32.37	32.37	0.005
	32.23		32.38		31.64		32.37		
3	32.41	32.43	32.60	32.59	32.36	32.40	32.43	32.43	0.005
	32.46		32.57		32.45		32.43		
4	32.15	32.15	32.42	32.44	32.05	32.08	32.16	32.14	0.019
	32.16		32.45		32.11		32.12		
5	31.95	31.97	32.33	32.36	31.94	31.93	31.98	32.01	0.030
	31.98		32.40		31.92		32.05		
6	32.01	32.02	32.32	32.30	31.96	31.95	32.01	32.03	0.018
	32.04		32.28		31.95		32.06		
7	31.70	31.71	32.22	32.09	31.79	31.73	31.83	31.81	0.024
	31.71		31.97		31.67		31.80		
8	31.51	31.52	31.91	31.91	31.59	31.55	31.64	31.63	0.024
	31.54		31.92		31.51		31.61		
9	31.11	31.12	31.54	31.52	31.26	31.21	31.38	31.36	0.023
	31.13		31.50		31.17		31.34		
10	31.25	31.26	31.69	31.66	31.36	31.32	31.44	31.42	0.024
	31.27		31.64		31.28		31.41		

Table 5.6 lists the PSNR obtained in each view and the perceived quality with two views achieved by the proposed proposed scheme when the channel bandwidth is 20MB/S. ‘V-PSNR’ in the table represents the PSNR obtained in each views and ‘P-PSNR’ indicates the perceived quality which is calculated based on (5.34). The variance of perceived quality among 4 stereo videos is also calculated in each GOP. Table 5.7 and Table 5.8 show the sets of results obtained from frame-compatible format with equal rate distribution and from full-resolution statistical multiplexing framework, respectively. It can be seen that the proposed scheme carried out symmetric bit allocation between views in this case. Based on the subjective viewing tests in [101], two views with similar quality achieves better overall perceptual quality if the JNT cannot be obtained by either view. Frame-compatible framework obtains higher PSNR and perceptual quality in sequence “Fountain” than other sequences. This is because the bandwidth is equally distributed among videos in frame-compatible framework, and for low complexity video such as “Fountain”, much higher quality can be obtained with

TABLE 5.7: PSNR in individual view, perceived quality, and variance of perceived quality for 4 stereo videos achieved by frame-compatible framework(Channel bandwidth = 20MB/S)

GOP Index	Street		Fountain		Lake		Rollerskate		P-Quality Variance
	V-PSNR	P-Quality	V-PSNR	P-Quality	V-PSNR	P-Quality	V-PSNR	P-Quality	
1	31.00	30.82	34.02	34.03	31.02	30.95	31.01	30.72	1.929
	30.63		34.03		30.87		30.42		
2	30.76	30.55	34.13	34.13	31.21	31.12	31.78	31.50	1.884
	30.35		33.99		31.02		31.21		
3	30.36	30.19	34.47	34.47	31.25	31.13	31.83	31.53	2.561
	30.02		34.31		31.01		31.23		
4	29.79	29.63	34.57	34.57	30.78	30.68	31.81	31.52	3.382
	29.48		34.43		30.58		31.24		
5	29.39	29.17	34.61	34.61	30.74	30.66	31.81	31.52	3.950
	28.95		34.49		30.57		31.23		
6	28.71	28.54	34.59	34.59	30.86	30.78	32.20	31.87	4.742
	28.37		34.48		30.70		31.54		
7	28.10	27.93	34.38	34.38	32.74	32.68	31.23	31.52	5.588
	27.76		34.25		32.62		31.81		
8	27.82	27.61	34.47	34.47	30.63	30.55	32.01	31.67	6.033
	27.41		34.35		30.47		31.33		
9	27.82	27.61	34.44	34.44	30.15	30.10	32.02	31.68	6.131
	27.41		34.30		30.06		31.35		
10	27.25	27.23	34.63	34.63	30.35	30.32	31.86	31.56	7.037
	27.20		34.37		30.29		31.27		

TABLE 5.8: PSNR in individual view, perceived quality, and variance of perceived quality for 4 stereo videos achieved by full-resolution statistical multiplexing framework(Channel bandwidth = 20MB/S)

GOP Index	Street		Fountain		Lake		Rollerskate		P-Quality Variance
	V-PSNR	P-Quality	V-PSNR	P-Quality	V-PSNR	P-Quality	V-PSNR	P-Quality	
1	34.33	34.33	34.99	34.99	31.64	31.80	35.02	35.68	2.141
	34.13		34.80		31.97		35.68		
2	33.60	33.76	34.57	34.92	32.78	32.93	34.69	34.99	0.738
	33.76		34.92		33.08		34.99		
3	33.68	33.85	34.95	35.16	33.22	33.48	34.66	34.88	0.484
	33.85		35.16		33.48		34.88		
4	33.36	33.36	35.05	35.29	33.02	32.84	34.47	34.88	1.047
	33.21		35.29		32.65		34.88		
5	33.38	33.16	34.58	34.94	33.21	33.33	33.57	34.17	0.502
	32.95		34.94		33.33		34.17		
6	32.29	32.42	34.53	35.04	33.27	33.57	34.47	34.88	1.127
	32.56		35.04		33.57		34.88		
7	31.88	32.01	34.47	34.87	33.72	33.79	33.80	34.19	1.123
	32.13		34.87		33.79		34.19		
8	31.88	31.85	34.62	34.95	32.79	32.92	34.02	34.02	1.349
	31.82		34.95		33.06		33.97		
9	31.49	31.52	34.66	34.66	32.28	32.30	34.27	34.57	1.912
	31.54		34.37		32.31		34.57		
10	31.53	31.53	35.12	35.12	32.65	32.64	33.71	34.15	1.902
	31.53		34.96		32.63		34.15		

TABLE 5.9: PSNR in individual view, overall perceived quality, and variance of perceived quality for 4 stereo videos achieved by proposed scheme (Channel bandwidth = 40MB/S)

GOP Index	Street		Fountain		Lake		Rollerskate		P-Quality Variance
	V-PSNR	P-Quality	V-PSNR	P-Quality	V-PSNR	P-Quality	V-PSNR	P-Quality	
1	37.67	37.67	37.55	37.55	37.45	37.45	37.70	37.70	0.010
	33.06		33.13		33.27		33.11		
2	37.40	37.40	37.56	37.56	37.59	37.59	37.59	37.59	0.007
	33.07		33.11		33.09		33.23		
3	37.50	37.50	37.74	37.74	37.64	37.64	37.66	37.66	0.008
	33.07		33.11		33.09		33.23		
4	37.10	37.10	37.20	37.20	37.04	37.04	37.35	37.35	0.013
	33.05		33.11		33.13		33.11		
5	36.87	36.87	36.96	36.96	37.01	37.01	36.91	36.91	0.003
	33.08		33.35		33.10		33.11		
6	36.80	36.80	37.05	37.05	37.09	37.09	36.91	36.91	0.013
	33.08		33.35		33.99		33.11		
7	36.50	36.50	36.60	36.60	36.74	36.74	36.75	36.75	0.010
	33.14		33.20		33.03		33.11		
8	36.21	36.21	36.17	36.17	36.33	36.33	36.24	36.24	0.004
	33.16		33.18		33.05		33.08		
9	35.44	35.44	35.65	35.65	35.84	35.84	35.68	33.51	0.020
	33.25		33.17		33.00		33.14		
10	35.62	35.62	35.85	35.85	36.00	36.00	35.87	35.87	0.019
	33.23		33.19		33.07		33.09		

the same amount of bits allocated. Statistical multiplexing framework, on the other hand, allocates channel bandwidth according to relative video complexity. However, the results turn to be highly dependent on the accuracy of the complexity measure. As comparing the results in Table 5.6 - 5.8, it can be seen that the variance of perceived quality is much smaller in the proposed scheme which proves that the proposed scheme achieves the objective of joint rate allocation, i.e. better inter-program fairness.

Experiments were conducted to further test the performance of the proposed scheme and frame-compatible scheme when the channel bandwidth is increased to 40 MB/S. The results are shown in Table 5.9 for the proposed scheme, Table 5.10 for frame-compatible framework and Table 5.11 for full-resolution framework. Under this higher channel bandwidth, it can be seen that the proposed scheme automatically adapts to asymmetric inter-view rate allocation to achieve higher perceived quality with high coding efficiency. The quality of auxiliary view is kept above JNT while more rates are allocated to reference view to improve the perceived

TABLE 5.10: PSNR in individual view, overall perceived quality, and variance of perceived quality for 4 stereo videos achieved by frame-compatible framework (Channel bandwidth = 40MB/S)

GOP Index	Street		Fountain		Lake		Rollerskate		P-Quality Variance
	V-PSNR	P-Quality	V-PSNR	P-Quality	V-PSNR	P-Quality	V-PSNR	P-Quality	
1	31.63	31.43	35.33	35.39	32.79	32.67	33.25	32.88	2.071
	31.22		35.39		32.55		32.51		
2	31.40	31.18	35.41	35.41	32.91	32.78	32.96	32.62	2.328
	30.96		35.28		32.66		32.27		
3	30.95	30.77	35.72	35.72	32.96	32.79	32.84	32.49	3.168
	30.59		35.58		32.63		32.14		
4	30.40	30.24	35.70	35.70	32.69	32.55	32.87	32.53	3.777
	30.08		35.61		32.41		32.19		
5	30.02	29.78	35.77	35.77	32.65	32.53	32.79	32.45	4.495
	29.55		35.67		32.41		32.12		
6	29.35	29.16	35.75	35.75	32.67	32.66	32.20	31.87	5.507
	28.97		35.67		32.65		31.54		
7	28.78	28.60	35.56	35.56	32.74	32.68	32.86	32.50	6.138
	28.41		35.46		32.62		32.14		
8	28.47	28.25	35.59	35.59	32.41	32.30	32.83	32.44	6.796
	28.02		35.49		32.19		32.05		
9	27.35	27.27	35.64	35.64	32.02	31.95	32.85	32.46	8.944
	27.18		35.53		31.88		32.07		
10	27.85	27.83	35.78	35.78	32.28	32.24	32.73	32.39	7.968
	27.80		35.51		32.20		32.05		

TABLE 5.11: PSNR in individual view, overall perceived quality, and variance of perceived quality for 4 stereo videos achieved by full-resolution statistical multiplexing framework (Channel bandwidth = 40MB/S)

GOP Index	Street		Fountain		Lake		Rollerskate		P-Quality Variance
	V-PSNR	P-Quality	V-PSNR	P-Quality	V-PSNR	P-Quality	V-PSNR	P-Quality	
1	37.59	37.59	38.02	38.13	35.14	35.26	39.05	39.69	2.524
	37.28		38.13		35.26		39.69		
2	36.38	36.53	37.46	37.94	36.47	36.89	38.41	38.86	0.836
	36.53		37.94		36.89		38.86		
3	36.35	36.78	38.22	38.22	37.19	37.52	37.82	38.58	0.475
	36.78		37.60		37.52		38.58		
4	36.01	36.12	38.22	38.22	36.26	36.64	38.55	38.55	1.049
	36.12		37.67		36.64		38.17		
5	36.16	36.16	37.45	37.64	37.34	37.81	36.80	37.80	0.477
	35.64		37.64		37.81		37.80		
6	35.41	35.92	37.59	37.59	37.58	37.82	37.55	37.99	0.686
	35.92		37.21		37.82		37.99		
7	34.97	35.16	37.10	37.50	37.99	38.06	37.03	37.03	1.189
	35.16		37.50		38.06		36.38		
8	34.69	34.84	37.42	37.72	36.76	37.14	37.19	37.70	1.398
	34.84		37.72		37.14		37.70		
9	34.27	34.27	37.35	37.35	35.98	36.08	37.57	37.92	1.962
	34.20		36.96		36.08		37.92		
10	34.60	34.61	37.42	37.76	36.28	36.31	37.20	37.37	1.486
	34.61		37.76		36.31		37.37		

quality. Frame-compatible framework gets much lower perceived quality in all videos in this set of results. It was observed in the experiment that, the PSNR for intermediate side-by-side frame-compatible sequences are increased with the higher channel bandwidth. However, the intermediate side-by-side frame-compatible sequence must be rearranged and upsampled to form two separate stereo views. The lost of spatial information has limited the quality increment on two stereo views of each video in frame-compatible framework. Full-resolution framework does not include the process of rearranging or upsampling and thus both PSNR of each video or perceived quality of each video increase with higher channel bandwidth. But similar as previous set of results, both frame-compatible framework and full-resolution framework get much bigger quality variation among videos as demonstrated by the high variance obtained in the results.

The increase of computational complexity in the proposed joint rate allocation scheme is due to the process of finding the optimal rates allocated to two views of each video, which is not required in the other two schemes. The adaptive inter-view rate allocation algorithm was implemented with C language on an Intel Core 2 Duo CPU @ 2.8 GHz and the algorithm takes up to 50 ms to solve the rate allocation problem for one GOP. Thus the increase of computational complexity in the proposed scheme can be ignored.

5.3.6 Summary

A scalable stereo video coding scheme that inherits some of the features from SVC has been proposed in this section. The reference view of stereo video is independently coded to provide backward compatibility while the auxiliary view is coded using the reference view as base layer to remove inter-view redundancy. The scalable format makes the rate control of each video as easy as bitstream truncation. For broadcast applications with multiple stereo

videos, the minimum video quality fluctuation should be achieved when switching from one video program to another. Targeting for this purpose, a joint bit allocation scheme has been proposed to dynamically distribute channel bandwidth among stereo videos. An adaptive inter-view rate allocation scheme has also been developed based on the characteristics of human visual system. Experimental results have shown that the proposed joint bit allocation scheme together with the proposed scalable stereo coding scheme achieve higher perceived quality in each video and smaller quality variation among all videos as compared with current technology used for 3DTV.

Chapter 6

Conclusions and Future Works

In this chapter, the main algorithm developments and results are summarized in Section 6.1, and some directions for future research are presented in Section 6.2.

6.1 Conclusions

With the significant progress in video coding technologies together with the rapid developments of network infrastructures, the demands for digital video applications are rapidly increasing and the expectation of the quality for these services is becoming higher. The ultimate objective in video coding is to provide the best video quality at the receiver end, given the constraint of certain network conditions. The bit allocation scheme is employed in many applications to distribute the limited bit budget among the video data. Though bit allocation is usually considered as an optimization problem, it is formulated differently for different applications with different types of video data. In this thesis, three types of application scenarios associated with three different bit allocation problems are studied. The contributions in the three types of application scenarios are:

- "One Video Source with One User"

Bit allocation in a video encoded for the requirement of a single use is commonly referred to as rate control. In Chapter 3, an adaptive scheme was proposed to perform adaptive GOP level bit allocation in rate control. The GOP size was first determined by a new dynamic threshold based keyframe detection scheme. The dynamic threshold is content adaptive and it is related to the resolution, QP as well as the statistical properties that are generated during the encoding of the video. Frames with either abrupt scene change or intensive motion movement in the video are detected and coded as a keyframe. Based on the keyframe detection scheme, a new method to properly determine the optimal GOP sizes has been proposed for rate control and a good trade off between coding efficiency and random access are obtained.

- "One Video Source to Multiple Users"

Scalable video coding (SVC) provides a scalable scheme for "one-source-multiple-use" in heterogeneous network. The layered structure of SVC allows one single bit stream to serve multiple users with different decoding and display capabilities and/or network channels with different bandwidths. In Chapter 4, three schemes were proposed to bring efficient implicit and cross-layer bit allocation in SVC. A combined SNR and spatial scalability scheme was proposed to provide SVC encoder the flexibility to optimize some specific layers that more customers are interested. A new concept of ROI was proposed and the weighting factors associated with each layer is decided by the customer composition in the organization. The method of Lagrangian optimization was employed for implicit bit allocation between motion and residual on layers so that the overall system performance can be maximized based on customer composition. A motion refinement scheme was then proposed to provide finer scalable granularity of the specific layer more customers are interested if the scalable bit rate range from the base layer to the specific layer is wide. An improved inter-layer prediction scheme

was proposed as the third scheme in the Chapter to handle problems arise from unbalanced cross layer bit allocation. The inter-layer prediction with two reference layers enhanced the coding efficiency of layers at high resolution with a better removal of redundancy. The new concept of auxiliary layer that can work with the new inter-layer prediction scheme was also introduced to solve existing unbalanced bit allocation.

- "Multiple Video Source to One User"

Multiple videos are available for a user to browse at any time instance in broadcast applications. In Chapter 5, joint bit allocation schemes have been proposed for multiple conventional videos and multiple stereo videos in broadcast applications. A new complexity measure that incorporates the characteristics of Human Visual System (HVS) was introduced to allocate minimum guaranteed channel bandwidth according to relative complexity of videos. The proposed video adaptation scheme in conventional broadcast applications not only minimizes the quality variations between videos, but also keeps the quality fluctuation within each video small. As 3D video is seen as new type of media for next generation broadcast and streaming applications, a scalable stereo coding format was proposed to make the rate adaptation of each stereo video as easy as bitstream truncation. By making use of the feature that human eyes can tolerate disparate qualities in two different views under some conditions, the proposed adaptive inter-view rate allocation scheme automatically switch between symmetric and asymmetric inter-view rate allocation according to network condition to provide better perceptual quality in each video. Together with the proposed joint rate allocation scheme, smaller quality variation among all videos is achieved as compared with current technology used for 3DTV.

6.2 Future Work

In addition to the bit allocation techniques discussed in this thesis, some interesting topics related to the work that have done can be extended for future research.

- **"Multiple Video Sources to Multiple Users" -Videoconferencing**

As mentioned in Chapter 1, all the application scenarios can be generally categorized into four groups. The fourth group - "Multiple Video Sources to Multiple Users" is one of the future directions. A typical application example in this scenario is multi-party videoconferencing, in which multiple participants can be seen on multiple screens at once. The participants involved in videoconferencing may have different display resolutions, transmission bandwidth or network conditions. Videos captured by different participants can also be different in terms of resolution, quality and etc. Many research problems on how to efficiently encode multiple videos as well as how to allocate limited bandwidth among multiple videos to ensure a maximized overall quality should be studied in this kind of applications. In addition, considering the realtime feature of videoconferencing, realtime SVC optimization is another important topic.

- **Cross-layer design for resource allocation in heterogeneous wireless networks**

The bit allocation techniques studies in this thesis are only concerned about resource allocation in source coding on application layer. With the increasing demand for high quality applications, providing quality of service (QoS) guaranteed services is necessary for future wireless networks. Different types of applications require diverse and widely varying QoS guarantees for various types of offered traffic. Traditionally, different protocol layers in the networks are treated as separate entities for specific operations. To fully optimize the wireless

broadband networks, both the challenges from the physical medium and the QoS demands from the applications have to be taken into account. Therefore, considerable works on cross-layer design are desired in order to provide efficient end-to-end QoS support for scalable video transmission.

- **Quality metric for 3D videos**

In Chapter 5, a joint rate allocation scheme for was proposed 3D videos. The rate allocation scheme was designed with the objective to achieve higher perceived quality in each video and almost uniform quality among multiple videos. There is one important problem to be solved here, which is, how to accurately measure the perceptual quality of 3D video quantitatively. The objective quality metric is not only useful for quality evaluation but also needed to solve some RD optimization problems. PSNR is a quality metric that is frequently used and widely adopted for conventional 2D videos and images. However, there are more views available in 3D video which makes PSNR not an effective metric. In the work presented earlier, a simple objective metric for stereo video proposed in [118] was used but a more accurate and more sophisticated metric is desired. The following requirements should be taken into consideration when designing the new objective quality metric:

- The quality metric should be accurate enough to align with perceptual quality;
- The quality metric should be simple;
- The quality metric should be differentiable. This is important as the optimization problem can only be solved when the model is differentiable;
- The quality metric must take the characteristics of HVS into consideration.

Author's Publications

Journal Publications

1. **W. Yao**, L.-P. Chau and S. Rahardja, "Joint Rate Allocation for Statistical Multiplexing in Video Broadcast Applications", *IEEE Trans. Broadcast.*, vol. 38, no. 3, pp. 417 - 427, Sep. 2012.
2. **W. Yao**, S. Rahardja and Z. G. Li, "Dynamic threshold based keyframe detection and its application in rate control", *IET Image Processing*, vol. 6, no. 7, pp. 986 - 995, Oct. 2012
3. **W. Yao**, L.-P. Chau and S. Rahardja, "Joint Rate Allocation for Stereoscopic 3D Videos in Next-Generation Broadcast Applications", *IEEE Trans. Broadcast.*, Accepted.
4. **W. Yao**, Z. G. Li and S. Rahardja, "A Scalable Video Coding in A Nutshell", Synthesis Journal by The Information Technology Standards Committee (ITSC), pp. 89 - 108, 2008.

Conference Publications

1. **W. Yao**, L.-P. Chau and S. Rahardja, "Joint Rate Allocation for Statistical Multiplexing of SVC", *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 725 - 728, Sep. 2012.
2. **W. Yao**, Z. G. Li and S. Rahardja, "A New Rate Control Scheme with Adaptive GOP sizes", *Proc. International Conference on Embedded Systems and Intelligent Technology (ICE-SIT)*, Feb. 2011.
3. **W. Yao** and S. Rahardja, "Dynamic Threshold based Keyframe Detection", *Proc. IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 2137 - 2140, Jun. 2010.
4. Z. G. Li, **W. Yao**, S. L. Xie and S. Rahardja, "Robust Image Similarity Indices Via Intensity Mapping Functions and Image Partition", *Proc. Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, Dec. 2010.

5. Z. G. Li, **W. Yao**, S. L. Xie and S. Rahardja, "A Similarity Index in the Complex Wavelet Domain", *Proc. IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1644 - 1649, Jun. 2010.
6. **W. Yao**, Z. G. Li and S. Rahardja, "Balanced Inter-layer Prediction for Combined Coarse Granular Scalability and Spatial Scalability". *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1759 - 1762, May 2007.
7. Z. G. Li, **W. Yao** and S. Rahardja, "Motion Refined Medium Granular Scalability for Scalable Video Coding". *Proc. Visual Communication and Image Processing (VCIP)*, Jan. 2007.
8. Z. G. Li, **W. Yao**, S. Rahardja and S. L. Xie, "New Framework for Encoder Optimization of Scalable Video Coding". *Proc. IEEE Workshop on Signal Processing Systems (SiPS)*, pp. 527 - 532, Oct. 2007.
9. Z. G. Li, X. Lin, S. Rahardja and **W. Yao**, "Customer Adaptive Combined SNR and Spatial Scalability". *Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1 - 4, Oct. 2005.

MPEG/JVT Proposals

1. Z. G. Li, **W. Yao** and S. Rahardja, S. L. Xie, "Encoder optimization of Scalable Video Coding". Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6.) JVT-X022, Geneva, Switzerland, Jul. 2007.
2. Z. G. Li, S. Rahardja, S. L. Xie and **W. Yao**, "Hypothetical Reference Decoder for Video Coding". Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6.) JVT-W020, San Jose, California, USA, Apr. 2007
3. **W. Yao**, Z. G. Li and S. Rahardja, "Balance Inter-layer prediction with two base layers". Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG (ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6.) JVT-V046, Marrakech, Morocco, Jan. 2007.

4. **W. Yao**, Z. G. Li and S. Rahardja, "Balance Inter-layer prediction". Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6. JVT-U021, Hangzhou, China, Oct. 2006.

Author's other publications not related to this thesis

1. **W. Yao**, Z. G. Li, S. Rahardja, S. S. Yao, J. H. Zheng, "Half-Quadratic Regularization based De-Noising for High Dynamic Range Image". *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp.1370 - 1373, Mar. 2010
2. **W. Yao** Z. G. Li and S. Rahardja, "Early Detection of All-Zero Block in H.264 With New Rate-Quantization Models". *Proc. IEEE International Symposium on Circuits and Systems (ISCAS)*. pp. 620 - 623, May 2008
3. Z. G. Li, **W. Yao**, S. L. Xie and S. Rahardja, "Price based Joint Source and Channel Coding for Real Time Video Applications over the Internet with Flow Rate Management", *Proc. IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1650 - 1656, Jun. 2010

Bibliography

- [1] *Video codec for audiovisual services at p x 64 kbit/s*" ITU-T Recommendation H.261, Version 1: Nov. 1990, Version 2: Mar. 1993.
- [2] *Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s - Part 2: Video*, ISO/IEC 11172-2 (MPEG-1 Video), Mar. 1993
- [3] *Information Technology-Generic coding of moving pictures and associated audio information - Part 2: Video*, ITU-T Recommendation H.262 and ISO/IEC 13818-2 (MPEG-2 Video), Nov. 1994
- [4] *Video coding for low bit rate communication*, ITU-T Recommendation H.263, Version 1: Nov. 1995, Version 2: Jan. 1998, Version 3: Nov. 2000
- [5] *Information Technology - Coding of audio-visual objects - Part 2: Visual*, ISO/IEC 14496-2 (MPEG-4 Visual), Version 1: Apr. 1999, Version 2: Feb. 2000, Version 3: May. 2004
- [6] *Information Technology-Coding of Audio-Visual Objects-Part 10: Advance Video Coding*, ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), Version 1: May. 2003, Version 2: Jan. 2004, Version 3: Sep. 2004, Version 4: Jul. 2005
- [7] T. Wiegand, G. J. Sullivan, G. Bjontegaard and A. Luthra "Overview of the H.264/AVC Video Coding Standard", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7 pp. 560-576, Jul. 2003.

- [8] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding: Extension of the H.264/AVC standard," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 17, no. 9, pp. 1103 - 1120, Sep. 2007.
- [9] *Information Technology-Coding of Audio-Visual Objects-Part 10: Advance Video Coding; Amendment 3 Scalable Video Coding*, ISO/IEC 14496-10:2005/AMD3.
- [10] Y. Wang, J. Ostermann and Y.-Q. Zhang, *Video Processing and Communications*, Upper Saddle River, New Jersey: Prentice Hall, 2002.
- [11] P. List, A. Joch, J. Lainema, G. Bjontegaard and M. Karczewicz, "Adaptive deblocking filter", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 614 - 619, Jul. 2003.
- [12] T. Wedi, "Motion compensation in H.264/AVC", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 577 - 586, Jul. 2003.
- [13] J. M. Boyce, "Weighted prediction in the H.264/MPEG AVC video coding standard", *Proc. Int. Symp. on Circuits and Systems*, pp. 789 - 792, 2004.
- [14] A. M. Tourapis, F. Wu, and S. Li, "Direct mode coding for bipredictive slices in the H.264 standard", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, pp. 119 - 126, Jan. 2005.
- [15] D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the h.264/avc video compression standard", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 620 - 636, Jul. 2003.
- [16] H. S. Malvar, A. Hallapuro, M. Karczewicz, and L. Kerofsky, "Low-complexity transform and quantization in h.264/avc", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 598 - 603, Jul. 2003.
- [17] M. Wien, H. Schwarz, and T. Oelbaum, "Performance analysis of SVC," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 17, no. 9, pp. 1194 - 1203, Sep. 2007.

- [18] H. Schwarz, T. Hinz, D. Marpe and T. Wiegand, "Constrained inter-layer prediction for single-loop decoding in spatial scalability", *Proc. of ICIP 2005*, Genova, Italy, 11-14 Sep. 2005
- [19] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971
- [20] K. R. Rao and P. Yip, *Discrete Cosine Transform*, Academic Press, 1990.
- [21] L. J. Lin and A. Ortega, "Bit-rate control using piecewise approximated ratedistortion characteristics", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 4, pp. 446-459, Aug. 1998.
- [22] H. M. Hang and J. J. Chen, "Source model for transform video coder and its application—Part I: fundamental theory", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 2, pp. 287-297, Apr. 1997.
- [23] J. J. Chen and H. M. Hang, "Source model for transform video coder and its application—Part II: variable frame rate coding", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 2, pp. 299-311, Apr. 1997.
- [24] T. Chiang and Y. Q. Zhang, "A new rate control scheme using quadratic rate distortion model", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 2, pp. 287- 311, Apr. 1997.
- [25] Z. He, Y. Kim and S. K. Mitra, "Low-delay rate control for DCT video coding via ρ -domain source modeling", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 8, pp. 928-940, Aug. 2001.
- [26] N. Kamaci, Y. Altunbasak and R. M. Mersereau, "Frame bit allocation for the H.264/AVC video coder via Cauchy-Density-Based rate and distortion models", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 8, pp 994-1006, Aug. 2005.
- [27] Y. Sun and I. Ahmad, "A robust and adaptive rate control algorithm for objectbased video coding", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 10, pp. 1167-1182, Oct. 2004.

- [28] X. Yang, W. Lin, Z. Lu, X. Lin, S. Rahardja, E. Ong and S. Yao, "Rate control for video-phone using local perceptual cues", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 4, pp. 496-506, Apr. 2005.
- [29] Y. Sun, I. Ahmad, D. Li and Y. Q. Zhang, "Region-based rate control and bit allocation for wireless video transmission", *IEEE Trans. Multimedia*, vol. 8, no. 1, pp. 1-10, Feb. 2006.
- [30] B. Xie and W. Zeng, "A sequence-based rate control framework for consistent quality real-time video", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 56-71, Jan. 2006.
- [31] M. Jiang and N. Ling, "Low-delay rate control for real-time H.264/AVC video coding", *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 467-477, Jun. 2006.
- [32] N. Changuel, B. Sayadi and M. Kieffer "Predictive control for efficient statistical multiplexing of digital video programs" *17th Int. Packet Video Workshop*, pp. 1 - 9, 2009.
- [33] Z. He, W. Zeng, and C. W. Chen, "Low-pass filtering of rate-distortion functions for quality smoothing in realtime video communication", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 8, pp. 973 - 981, 2005.
- [34] G. Valenzise, M. Tagliasacchi and S. Tubaro, "A smoothed, minimum distortion-variance rate control algorithm for multiplexed transcoded video sequences", *Proc. Int. Multimedia Conf.*, pp. 55 - 60, 2007.
- [35] "MPEG-2 Test Model 5," Doc. ISO/IEC JTC1/SC29 WG11/93-400, Apr. 1993.
- [36] J. R-Corbera, and S. Lei, "Rate control in DCT video voding for low-delay communication," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 9, no. 1, pp. 172 - 185, Feb. 1999.
- [37] H. J. Lee, T. H. Chiang, and Y.-Q. Zhang, "Scalable rate control for MPEG-4 video," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 10, pp. 878 - 894, Sep. 2000.
- [38] MPEG-4 Video Verification Model V18.0, Coding of Moving Pictures and Audio N3908, ISO/IEC, JTC1/SC29/WG11, Jan. 2001.

- [39] Z. G. Li, W. Gao, F. Pan, S. W. Ma, K. P. Lim, G. N. Feng, X. Lin, S. Rahardja, H. Q. Lu and Y. Lu, "Adaptive rate control for H.264", *J. of Visual Commun. and Image Representation*, vol. 17, no. 2, pp. 376-406, 2006.
- [40] J. Lee, I. Shin and H. W. Park, "Adaptive Intra-Frame Assignment and Bit-Rate Estimation for Variable GOP Length in H.264", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 10, pp. 1271 - 1279, 2006,
- [41] A. Y. Lan, A. G. Nguyen and J. N. Hwang, "Scene-context-dependent reference-frame replacement for MPEG video coding", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 3, pp 478-489, 1999.
- [42] Y. Yuan, D. Feng and Y. Zhong "A novel method of keyframe setting in video coding: fast adaptive dynamic keyframe selecting", *Proc. Int. Conf. on Computer Networks and Mobile Computing*, pp.428 - 431, Oct. 2003.
- [43] J.-R. Ding, and J.-F. Yang, "Adaptive group-of-picture and scene change detection methods based on existing H.264 advanced video coding information", *IET Image Processing*, vol. 2, no. 2, pp. 85-94, 2008
- [44] A. Dimou, O. Nemethova and M. Rupp, "Scene change detection for H.264 using dynamic threshold technique", *Proc. of 5th EURASIP Conf.*, 2005.
- [45] S. Youm and W. Kim, "Dynamic threshold method for scene change detection", *Proc. Int. Conf. on Multimedia & Expo*, vol. 2, pp. 337-340, Jul. 2003,
- [46] H. Li, G. Liu, Z. Zhang and Y. Li, "Adaptive scene-detection algorithm for VBR video stream", *IEEE Trans. Multimedia*, vol. 6, no. 4, pp. 624-633, 2004.
- [47] X. Wang and Z. Weng, "Scene abrupt change detection", *Proc. Canadian Conf. on Elect. and Comput. Eng.*, vol. 2, pp. 880-883, 2000.
- [48] K. W. Sze, K. M. Lam and G. Qiu, "Scene cut detection using the colored pattern appearance model", *Proc. Int. Conf. on Image Processing*, vol. 2, pp. 1017-1020, Sep. 2003.

- [49] P. K. Sahoo, S. Soltani, A. K. C. Wong and Y. C. Chen, "SA survey of thresholding techniques", *Proc. of Computer Vision, Graphics, and Image Processing*, vol. 41, pp. 233-260, 1988.
- [50] P. Usach-Molina, J. Sastre, V. Naranjo, L. Vergara and J. M. L. Munoz "Content-Based Dynamic Threshold Method for Real-Time Keyframe Selecting", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no.7 , pp.982 - 993, 2010.
- [51] 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. 472-487, 2003.
- [52] ITU-T/SG15, Video codec test model, TMN7, Nice, Feb. 1997.
- [53] J. Ribas-Corbera and S. Lei, "Rate control in DCT video coding for low-delay communications", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 1, pp. 172 - 185, Feb. 1999.
- [54] J. Ribas-Corbera and S. Lei, "A frame-layer bit allocation for H.263+", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 7, pp. 1154 - 1158, Oct. 2000.
- [55] H.264/MPEG-4 AVC Reference Software, Version 14, "<http://iphome.hhi.de/suehring/tml/download/>".
- [56] G. Sullivan, "Recommended simulation common conditions for H.26L coding efficiency experiments on low-resolution progressive-scan source material", ITU-T VCEG, 14th Meeting, VCEG-N81.doc, Sep. 2001 Santa Barbara, CA, USA
- [57] D. S. Taubman and M. W. Marcellin, JPEG2000 Image Compression, Fundamentals, Standards and Practice. Kulwer Academic Publishers, 2002.
- [58] H. C. Huang, C. N. Wang and T. H. Chiang, "A Robust Fine Granularity Scalability Using Trellis-Based Predictive Leak", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, no. 6 pp. 372-384, 2002.

- [59] F. Wu, S. P. Li, and Y. Q. Zhang, "A Framework for Efficient Progressive Fine Granularity Scalable Video Coding", *IEEE Trans. on Circuits and Systems for Video Technol.*, vol. 11, no. 3, pp. 332-344, Mar. 2001.
- [60] M. V. Schar "A hybrid temporal-SNR fine-granular scalability for internet video", *IEEE Trans. on Circuits and Systems for Video Technol.*, vol. 11, no. 3, pp. 318-331, Mar. 2001.
- [61] H. Schwarz, D. Marpe and T. Wiegand, "SNR-Scalable Extension of H.264/AVC", *Proc. of ICIP 2004*, Singapore, 24-27 Oct. 2004.
- [62] Z. G. Li, S. Rahardja, H. F. Sun, "Implicit Bit Allocation for Combined Coarse Granular Scalability and Spatial Scalability", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 12, pp. 1449-1459, 2006
- [63] Z. G. Li and S. Rahardja, "Customer Oriented SNR Scalability for Scalable Video Coding", *Proc. of VCIP 2005*, Beijing, P. R. China, 12-15 July. 2005
- [64] H. Schwarz, D. Marpe and T. Wiegand, "Inter-Layer Prediction of Motion and Residual Data", ISO/IEC JTC1/SC29/WG11 doc. M11043. Redmond, USA. 19-23 July. 2004.
- [65] H. Schwarz, D. Marpe and T. Wiegand, "Constraint inter-layer prediction for single-loop decoding in spatial scalability", *Proc. of ICIP 2005*, Genoa, Italy, vol.2, pp. 870-873, Sep. 2005.
- [66] Z. G. Li, S. Rahardja, X. Lin and W. Yao, "Customer adaptive combined SNR and spatial scalability scheme", *International Workshop on Multimedia Signal Processing (MMSP2005)*, Shanghai, 30 Oct. - 2 Nov., 2005.
- [67] I. Amonou, N. Cammas, S. Kervadec and S. Pateux, "Improving Inter-layer Prediction", ISO/IEC JTC1/SC29/WG11 and ITU-T SG 15 Q.6 JVT-T053, Klagenfurt, Austria, 15-21 July, 2006.
- [68] HHI SVC Software Version 1.0, <http://ftp.tnt.uni-hannover.de>.

- [69] L. Böröczky, A. Y. Ngai and E. F. Westermann, "Statistical multiplexing using MPEG-2 video encoders," *Int. Bus. Machines J. Res. Develop.*, vol. 43, no. 4, pp. 511-520, Jul. 1999
- [70] L. Wang and A. Vincent, "Joint rate control for multiprogram video coding", *IEEE Trans. Consum. Electron.*, vol. 42, no. 3, pp. 300 - 305, Aug. 1996.
- [71] J. Yang, X. Fang, and H. Xiong, "A joint rate control scheme for H.264 encoding of multiple video sequences", *IEEE Trans. Consum. Electron.*, vol. 51, no. 2, pp. 617 - 623, May 2005.
- [72] M. Perkins and D. Arnstein, "Statistical multiplexing of multiple MPEG-2 video programs in a single channel", *Soc. Motion Picture Television Engineers J.*, vol. 104, no. 9, pp. 596 - 599, 1995.
- [73] Z. He and D. O. Wu, "Linear rate control and optimum statistical multiplexing for H.264 video broadcast", *IEEE Trans. on Multimedia*, vol. 10, no. 7, pp. 1237 - 1249, Nov. 2008.
- [74] L. Böröczky, A. Y. Ngai, and E. F. Westermann, "Joint rate control with look-ahead for multiprogram video coding," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 10, no. 7, pp. 1159-1163, Oct. 2000.
- [75] E. N. Linzer and A. Wells, "Statistical multiplexed video encoding using preencoding a priori statistics and a priori and a posteriori statistics", U.S. Patent 6 094 457, Jul. 2000.
- [76] Z. G. Li, C. Zhu, F. Pan, G. N. Feng, X. K. Yang, S. Wu and N. Ling, "A Novel Joint Rate Control Scheme for the Coding of Multiple Real Time Video Programs," in *Proc. Distributed Computing Syst. Workshops.*, pp. 241-245, Nov. 2002.
- [77] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini and G.J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 688 - 703, Jul. 2003.

- [78] T. Paridaens, D. De Schrijver, W. De Neve, and R. Van de Walle, "XMLdriven bit-rate adaptation of SVC bit-streams", in *Proc. Int. Workshop Image Anal. Multimedia Interactive Services*, 2007, pp. 49 - 52.
- [79] S. K. Srinivasan, J. Vahabzadeh-Hagh and M. Reisslein, "The Effects of Priority Levels and Buffering on the Statistical Multiplexing of Single-Layer H.264/AVC and SVC Encoded Video Streams," *IEEE Trans. Broadcast.*, vol. 56, no. 3, pp. 281-287, Sep. 2010.
- [80] C. Heyaime-Duvergé, V. K. Prabhu, "Statistical Multiplexing of Upstream Transmissions in DOCSIS Cable Networks," *IEEE Trans. Broadcast.*, vol. 56, no. 3, pp. 296-310, Sep. 2010.
- [81] H. Lahdili, H. Najaf-Zadeh and L. Thibault, "Statistical Multiplexing for Digital Audio Broadcasting Applications," *IEEE Trans. Broadcast.*, vol. 56, no. 1, pp. 19-27, Mar. 2010.
- [82] G. Van der Auwera and M. Reisslein, "Implications of Smoothing on Statistical Multiplexing of H.264/AVC and SVC Video Streams," *IEEE Trans. Broadcast.*, vol. 55, no. 3, pp. 541-558, Sep. 2009.
- [83] X. Yang and N. Ling, "Statistical multiplexing based on mpeg-4 fine granularity scalability coding", *J. VLSI Signal Process. Syst.*, vol. 42, no. 1, pp. 69 - 77, 2006.
- [84] M. Jacobs, J. Barbarien, S. Tondeur, R. Van de Walle, T. Paridaens and P. Schelkens, "Statistical multiplexing using SVC", in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast.*, 2008, pp. 1 - 6.
- [85] Y. Wang, L.-P. Chau and K.-H. Yap, "Joint rate allocation for multiprogram video coding using FGS," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 20, no. 6, pp. 829-837, June. 2010.
- [86] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
- [87] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Measurement to Structural Similarity", *IEEE Trans. on Image Process.*, vol. 13, no.4, pp. 600-613, Apr. 2004.

- [88] S. Chikkerur, V. Sundaram, M. Reisslein and L. J. Karam, "Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165-182, June 2011.
- [89] S. S. Hemami, A. R. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design", *Signal Processing: Image Communication*, vol. 25, no.7, pp. 469-481, Aug. 2010.
- [90] K. Seshadrinathan, T. Soundararajan, A. C. Bovik and L. K. Cormack, "Study of Subjective and Objective Quality Assessment of Video," *IEEE Trans. on Image Process.*, vol. 19, no. 6, pp. 1427-1441, June 2010.
- [91] A. C. Brooks, X. Zhao and T. N. Pappas, "Structural Similarity Quality Metrics in a Coding Context: Exploring the Space of Realistic Distortions", *IEEE Trans. on Image Process.*, vol. 17, no.8, pp. 1261 - 1273, Aug. 2008.
- [92] Y.-H. Huang, T.-S. Ou, P.-Y. Su and H. H. Chen, "Perceptual Rate-Distortion Optimization Using Structural Similarity Index as Quality Metric", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1614 - 1624, Nov. 2010.
- [93] Z. G. Li, W. Yao, S. L. Xie and S. Rahardja, "Robust image similarity indices via intensity mapping functions and image partition", in *Proc. Asia-Pacific Signal and Information Processing Association Annu. Summit and Conf.*, Dec. 2010.
- [94] Z. G. Li, Z. J. Zhu and S. Rahardja, "Structural similarity indices for high dynamic range imaging" in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1145 - 1148, May 2011.
- [95] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 3, pp. 301 - 317, Mar. 2001.
- [96] W. J. Kim, J. W. Yi, and S. D. Kim, "A bit allocation method based on picture activity for still image coding", *IEEE Trans. Image Process.*, vol. 8, no. 7, pp. 974 - 977, Jul. 1999.

- [97] J. Sun, W. Gao, D. Zhao, and W. Li, "On rate-distortion modeling and extraction of H.264/SVC fine-granular scalable video", *IEEE Trans. Circuit Syst. Video Technol.*, vol. 19, no. 3, pp. 323 - 336, Mar. 2009.
- [98] J. Reichel, H. Schwarz and M. Wien, *Joint Scalable Video Model JSVM-6*, ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6 JVT-S202. Geveva, Switzerland, April, 2006.
- [99] *JSVM 9.8 Software*, ISO/IEC JTC 1/SC 29/WG 11 W8752, MPEG Committee, 2007.
- [100] A. Vetro, A. M. Tourapis, K. Muller and T. Chen, "3D-TV content storage and transmission", *IEEE Trans. Broadcast*, vol. 57, no. 2, pp. 384 - 394, Jun. 2011.
- [101] G. Saygili, C. G. Gurler and A. M. Tekalp, "Evaluation of asymmetric stereo video coding and rate scaling for adaptive 3D video streaming", *IEEE Trans. Broadcast*, vol. 57, no. 2, pp. 593 - 601, Jun. 2011.
- [102] DVB BlueBook A151, *Commercial Requirements for DVB 3D-TV*, Jul. 2010.
- [103] *HDMI Specification 1.4a*, HDMI Licensing, LLC., Mar. 2010.
- [104] *Text of ISO/IEC FDIS 23002-3 Representation of Auxiliary Video and Supplemental Information*, ISO/IEC JTC1/SC29/WG11, Marrakesh, Morocco, Doc. N8768, Jan. 2007.
- [105] A. M. Tourapis, P. Pahalawatta, A. Leontaris, Y. He, Y. Ye, K. Stec and W. Husak, *A frame compatible system for 3D delivery*, ISO/IEC JTC1/SC29/WG11, Doc. M17925, Jul. 2010.
- [106] W.-S. Kim, A. Ortega, P. Lai, D. Tian and C. Gomila, "Depth map coding with distortion estimation of rendered view", *Proc. SPIE Visual Inf. Process. Commun.*, vol. 7543, 2010.
- [107] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder and R. Szeliski, "High-quality video view interpolation using a layered representation", *ACM SIGGRAPH and ACM Trans. Graphics*, vol. 23, no.3, pp. 600 - 608, Aug. 2004.

- [108] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff and T. Wiegand, "View synthesis for advanced 3D video systems", *EURASIP J. Image Video Process.*, Special Issue on 3D Image Video Process., vol. 2008, p. 11, 2008.
- [109] M. G. Perkins, "Data compression of stereo pairs", *IEEE Trans. Commun.*, vol. 40, no. 4, pp. 684 - 696, Apr. 1992.
- [110] L. Stelmach and W. J. Tam, "Stereoscopic image coding: Effect of disparate image-quality in left- and right-eye views" *Signal Process.: Image Commun.*, vol. 14, pp. 111 - 117, 1998.
- [111] L. B. Stelmach, W. J. Tam, D. V. Meegan, A. Vincent and P. Corriveau, "Human perception of mismatched stereoscopic 3D input", *Proc. IEEE Int. Conf. Image Process.*, vol. 1, pp. 5 - 8, Sep. 2000.
- [112] L. M. J. Meesters, W. A. IJsselstein and P. J. H. Seuntings, "A survey of perceptual evaluations and requirements of three-dimensional TV", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, pp. 381 - 391, Mar. 2004.
- [113] L. B. Stelmach, W. J. Tam, D. V. Meegan and A. Vincent, "Stereo image quality: Effects of mixed spatio-temporal resolution", *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 2, pp. 188 - 193, Mar. 2000.
- [114] P. Aflaki, M. M. Hannuksela, J. Hakkinen, P. Lindroos and M. Gabbouj, "Subjective study on compressed asymmetric stereoscopic video", *Proc. IEEE Int. Conf. Image Process.*, pp. 4021 - 4024, Sep. 2010.
- [115] C. Fehn, P. Kauff, S. Cho, H. Kwon, N. Hur and J. Kim, "Asymmetric coding of stereoscopic video for transmission over T-DMB", *Proc. 3DTV Conf.*, pp. 1-4, May 2007. 4379449.
- [116] A. Aksay, C. Bilen, E. Kurutepe, T. Ozcelebi, G. B. Akar, M. R. Civanlar and A. M. Tekalp, "Temporal and spatial scaling for stereoscopic video compression", *Proc. 14th Eur. Signal Process. Conf.*, Sep. 2006.

- [117] W. J. Tam, L. B. Stelmach, F. Speranza and R. Renaud, "Crossswitching in asymmetrical coding for stereoscopic video" *Stereoscopic Displays and Virtual Reality Systems IX*, vol. 4660, pp. 95 - 104, 2002
- [118] G. Saygili, G. Gurler, and A. M. Tekalp, "3D display-dependent quality evaluation and rate allocation using scalable video coding", *Proc. IEEE Int. Conf. Image Process*, pp. 717 - 720, 2009
- [119] G. Saygili, G. Gurler, and A. M. Tekalp, "Quality assessment of asymmetric stereo video coding", *Proc. IEEE Int. Conf. Image Process.*, pp. 4009 - 4012, Sep. 2010,
- [120] N. Ozbek and A.M. Tekalp, "Unequal inter-view rate allocation using scalable stereo video coding and an objective stereo video quality measure," *Proc. IEEE Int. Conf. Multimedia and Expo*, pp. 1113 - 1116, 2008.
- [121] F. Lu, H. Wang, X. Ji, and G. Er, "Quality assessment of 3D asymmetric view coding using spatial frequency dominance model", *Proc. 3DTV Conference*, pp. 1 - 4, May 2009.
- [122] T. F. Edgar and D. M. Himmelblau, *Optimization of Chemical Processes*. New York: McGraw-Hill, 1988.