

# **Compression and error robustness performance improvement of a scalable multi view video coding frame work**

**THESIS**

Submitted in partial fulfillment  
of the requirements for the degree of  
**DOCTOR OF PHILOSOPHY**

by

**PALANIVEL GURUVAREDDIAR**

Under the Supervision of

**Dr. Biju K. Joseph**



**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE**

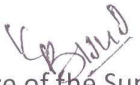
**PILANI (RAJASTHAN) INDIA**

**2016**

**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI**

**CERTIFICATE**

This is to certify that the thesis entitled **Compression and error robustness performance improvement of a scalable multi view video coding frame work** and submitted by **Palanivel Guruvareddiar** ID No **2008PHXF436P** for award of Ph.D. of the Institute embodies original work done by him/her under my supervision.

  
Signature of the Supervisor  
Name in capital letters **Biju K. JOSEPH**  
Designation **Manager**

Date: **Dec 02, 2016**

## ABSTRACT

The advanced video coding based multiview video coding (H.264 MVC) is the current state-of-the-art compression scheme for 3D video and is adopted in a number of products such as 3D Blu-ray, Sony creative software, Intel media SDK, Imagination technologies' powerVR VXD392 IP core. It is well poised to be used as a signal of information for new 'content - rich' visual media such as 3DTV, free view point TV and 3D video conferencing. H.264 MVC provides only temporal and view scalability to support different frame rates and legacy 2D video devices respectively. However, in order to support a number of 3D capable devices over heterogeneous networks spatial and bit rate scalability will be preferred. The aim of the thesis is to introduce a novel scalable multiview video compression scheme which provides temporal, view, spatial and bit rate scalabilities with backward compatibility to existing 2D and 3D video devices with emphasis on single loop decoding and parallel decoding of layers.

Scalable multiview video compression schemes based on wavelets and hybrid coding based methods are reviewed in the thesis. Wavelet-based scalable multi view video coding method has a number of limitations as mentioned below,

- Lack of backward compatibility with existing 2D and 3D legacy end points
- Operational mismatch between encoder and decoder
- Requires good amount of memory
- Increased system latency
- Requires multiple loop decoding

On the other hand, hybrid coding based scalable compression, at least for 2D video is not a new technology. Many earlier standards such as MPEG-2/H.262, H.263 and MPEG-4 have included tools to provide several important scalabilities, though these profiles have rarely been used. The major reason for the failure of these profiles is that the decoding complexity was fairly high due to multiple loop decoding and these profiles usually come along with loss in compression efficiency. In this thesis emphasis has been given to single loop decoding wherein the decoding is limited to the layer for which the decoder is configured and only parsing process is required for reference layers. Also more emphasis is placed on the parallel coding of layers and simulation results show that compression efficiency of the proposed scheme is better than the simulcast methods. The thesis is based on the H.264 multi-view video coding specification, which is the current state-of-the-art in the hybrid coding based compression schemes for multi-view video and some of the introduced coding tools have been derived from H.264 scalable video coding specification for 2D video.

The application areas of the thesis include 3DTV, immersive video conferencing and this thesis present methods targeted towards both first and second phases of 3DTV deployments as summarized below,

- A method to improve the existing “frame compatible format” has been proposed, where the left and right views of the stereoscopic video frame will be packed into two separate tiles of the video frame. The resultant frame will be encoded using HEVC compression scheme and additionally the loop filters will be turned off across the tile boundaries. This is targeted towards the first phase of 3DTV deployments where the existing 2D infrastructure will be retained and the proposed method improves the coding speed by approximately twice while maintaining the same rate-distortion performance.
- Following methods targeted toward the second phase of 3DTV deployments have been proposed. The second phase is termed as the “service compatible” and will be realized using the newly adopted multi-view video compression schemes and infrastructure.
  - Methods have been suggested in this thesis to improve the existing temporal scalability of H.264 multi-view video coding specification, by which the temporal identifier information is used in the reference picture list construction process in addition to the sub-bitstream extraction process.
  - A disposable view components based temporally scalable and view scalable prediction structure has been presented. This prediction structure provides improved compression efficiency compared to the existing prediction structures and also enhances the ability of the media aware network server to create sub sequences based on the target device’s capability and network in use.
  - The thesis presents the spatial scalability and coarse grain bit rate scalability using the layered coding approach, where the inter-layer redundancies are exploited along with the intra-layer redundancies. The methods presented in the thesis have been benchmarked against the simulcast and simulation results show that superior compression efficiency could be achieved by using the proposed schemes.
  - The thesis also presents the error robustness provided by the proposed architecture. Simulation results show that the error robustness method outlined in the thesis could be used to combat the reduction in transmission bandwidth using selective dropping of video packets. Experimental results show that there is no perceivable difference in video quality.

The various coding tools presented in the thesis have been carefully evaluated from a number of different perspectives such as compression efficiency, backward compatibility, stereoscopic 3D support, re-use of existing infrastructure, single loop decoding, parallel decoding of layers and support for various scalability options. From the analysis and simulation results it could be seen that the method proposed in this thesis could be used for the seamless inter-operability of various multi-view video capable devices, including the existing legacy 2D and/or 3D devices, connected with each other over a variety of heterogeneous networks.

## **Preface**

The research presented in this thesis has been carried out during the years 2009-2016 at Tata Elxsi Limited. During the preparation of the thesis, the author worked with several Tata Elxsi business units, including multimedia, systems and transportation. To start with, I would like to express my deepest gratitude to my supervisor Dr.Biju.K.Joseph for encouragement, ideas and constant guidance throughout the years as well as careful review of the thesis and publications.

I wish to express my thanks to the following authors for granting permission to use figures; Anthony Vetro, Sehoon Yea, Matthias Zwicker, Wojciech Matusik and Hanspeter Pfister for Fig.3.1. Anthony Vetro, Thomas Wiegand and Gary Sullivan for Fig.3.3. Philipp Merkle, Aljoscha Smolic, Karsten Müller and Thomas Wiegand for Fig.3.4. C. Hewage, S. Warrall, S. Dogan and A. M. Kondo for Fig.8.2, Fig.8.3. U. Fecker., J. Seiler., and A. Kaup for Fig.8.4. I also wish to express my deepest gratitude to doctoral advisory committee members of BITS for the constant guidance and support over the years.

I would like to thank the reviewers of the thesis, Dr. Pawan Ajmera and Dr.K.K. Gupta for their valuable review comments.

I owe many thanks to my former supervisors Dr. Sachin Gengaje and Dr.Vineet Gupta for the fruitful discussions and in formulating the research problems. During the years I worked with some of the great colleagues at Tata Elxsi Limited and I would like to thank Mr. Krishna Seshadri, Mr. Rajeev Kumar, Mr. Vadivel Shanmugam and Mr. Sandeep Kumar Soni for reviewing some of my publications and providing fruitful review comments.

I owe a lot to my father Gurusvareddiar for seeding the thoughts to become a doctor when I was in the early years of my bachelors in engineering. I also wish to express my warmest thanks to mother Sivakami, my brother Gurunathan and my wife Amudha, my kids Guru Vikranth and Guru Darshini for providing the much needed encouragement to complete the thesis.

Bangalore, December 2016

Palanivel Gurusvareddiar

## Table of Contents

Chapter 1 .....	13
Chapter 2 .....	20
2.1. Spatial interleaving using Frame Packing Arrangement .....	20
2.2. Temporal interleaving using SVC .....	22
2.3. Wavelet based methods .....	24
2.4. Hybrid coding based scalable 3D .....	26
Chapter 3 .....	31
3.1. A brief overview of the H.264/AVC standard.....	31
3.1.1. Profile and Level .....	32
3.1.2. Encoder operation.....	33
3.1.3. Multiple reference pictures and Decoded Picture Buffer .....	35
3.1.4. Memory management process .....	36
3.1.5. Prediction structures .....	38
3.2. A brief overview of the H.264 MVC standard .....	38
3.2.1. Profiles and level .....	39
3.2.2. MVC Prediction Structures .....	40
3.2.3. Encoder operation.....	42
Chapter 4 .....	43
4.1. Layered architecture .....	43
4.2. Extraction and adaptation of bitstreams .....	45
4.3. Software overview.....	47
4.4. Test procedure and Test sequences .....	49
Chapter 5 .....	51
5.1. Prediction structures for H.264 MVC.....	51
5.1.1. Disposable view components based hierarchical coding.....	51
5.1.2. Temporal identifier based hierarchical coding .....	53
5.1.3. Proposed prediction structure .....	54
5.1.4. Simulation setup and results .....	55
5.2. Improved temporal scalability for H.264 MVC .....	57
5.2.1. Reference picture list construction process .....	57
5.2.2. Modified list construction process.....	59

5.2.3. Simulation setup and results .....	61
Chapter 6 .....	63
6.1. Inter-layer prediction .....	63
6.1.1. Inter-layer intra prediction .....	64
6.1.2. Inter-layer motion prediction .....	65
6.1.3. Inter-layer residual prediction .....	66
6.1.4. Disparity compensated prediction .....	66
6.1.5. Single loop decoding .....	67
6.1.6. Signaling mechanisms .....	67
6.2. Simulation setup and results .....	69
6.2.1. Comparison between H.264 SVC and HEVC Up-sampling filters .....	69
6.2.2. Evaluation for spatial scalability .....	70
6.2.3. Spatio-temporal scalability .....	76
Chapter 7 .....	81
7.1. Coarse grain (CGS) bit rate scalability .....	81
7.2. Medium grain bit rate scalability .....	86
7.3. CGS-temporal scalability .....	89
Chapter 8 .....	94
8.1. Need for error concealment .....	94
8.2. Error concealment schemes in H.264 MVC .....	95
8.2.1. Depth image based error-concealment schemes .....	95
8.2.2. Disparity vector based frame-loss concealment schemes .....	97
8.2.3. Disparity vector based slice-loss concealment schemes .....	97
8.2.4. Concealment using decoder side ME .....	99
8.2.5. Forward error correction based concealment schemes .....	99
8.3. Proposed error-concealment scheme using SMVC .....	99
8.3.1. Error concealment for CGS case .....	100
8.3.2. Error concealment for SS case .....	105
Chapter 9 .....	110
Bibliography .....	113
Biography .....	121

## List of publications

The thesis is written on the basis of the following publications.

[C1] Gurusvareddiar, Palanivel, and Biju K. Joseph. "Frame-Compatible Stereo 3D Services Using H. 264/AVC and HEVC." *Data Compression Conference (DCC), 2013*. IEEE, 2013.

[J1] Gurusvareddiar, Palanivel, and Biju K. Joseph. "Comparative study of frame-compatible stereo 3D services and a novel method for spatial interleaving using HEVC." *Journal of Visual Communication and Image Representation* 26 (2015): 200-209.

[J2] Gurusvareddiar, Palanivel, and Biju K. Joseph. "Efficient Prediction Structures for H. 264 Multi View Coding Using Temporal Scalability." *3D Research* 5.1 (2014): 1-9.

[P1] Gurusvareddiar, Palanivel; Biju K. Joseph, "METHOD OF MANAGING REFERENCE PICTURE LIST FOR A MULTI VIEW VIDEO SIGNAL" (filed to India Patent Office, 2014)



## List of Figures

Figure 1.1: Camera arrangements for capturing sequences with multiple camera system. Camera setups are: 1-D line (left) and 1-D arc (right).....	13
Figure 1.2: Overview of multi view video system .....	14
Figure 1.3: Immersive 3D conferencing system.....	16
Figure 2.1.a: SbS frame compatible format. ‘X’ denotes the samples from left view and ‘O’ denotes the samples from right view. ....	21
Figure 2.1.b: TaB frame compatible format. ‘X’ denotes the samples from left view and ‘O’ denotes the samples from right view. ....	21
Figure 2.2: Prediction structure with 2 reference frames / list for view scalability.....	23
Figure 2.3: SVC- TS prediction structure with 2 reference frames / list.....	23
Figure 2.4: SVC-TS prediction structure with 4 reference frames / list.....	24
Figure 2.5: R-D curves for various H.264 based methods for akko & kayo sequence.....	28
Figure 2.6: R-D curves for various H.264 based methods for crowd sequence .....	28
Figure 2.7: R-D curves for various HEVC based methods for akko & kayo sequence.....	29
Figure 2.8: R-D curves for various HEVC based methods for crowd sequence .....	29
Figure 3.1: High-level coding architecture of the H.264/AVC encoder [37] .....	34
Figure 3.2: Prediction structure with quantization level set corresponding to the layer. Base layer will be encoded using lowest QP. ....	38
Figure 3.3: An illustration of MVC profiles, consisting of the multi view High and Stereo High profiles [11].....	39
Figure 3.4: State-of-the art prediction structure for H.264 MVC [53] .....	41
Figure 4.1: Layered architecture of the scalable multi view coding encoder with two layers. ....	43
Figure 4.2: Packet arrangement of the output stream of the scalable multi view coding encoder. Base view of the base layer is AVC compatible 2D view and rest of all layers is MVC compliant.....	44
Figure 4.3: NAL header of H.264/MVC complaint stream. Solid box indicates the reserved bit.....	46
Figure 4.4: Packet arrangement of the SMVC scheme.....	46
Figure 4.5: Packet arrangement of the SMVC scheme with two spatial layers, one quality layer each with three temporal and three view layers.....	47
Figure 4.6: Overview of SMVC software using JM reference software .....	48
Figure 4.7: Encoding of multiple view videos using SMVC encoder .....	50
Figure 4.8: Extraction of required packets using JMVC bitstream extractor .....	50
Figure 4.9: Decoding of SMVC bitstream using SMVC decoder (JM reference software).....	50
Figure 5.1: Disposable B frames based hierarchical coding. Dashed lines indicate inter-view prediction and solid lines represents temporal prediction. The top-box represents view components from right view camera and the bottom box depicts the view components from left view camera. ....	52
Figure 5.2: Temporal id based hierarchical coding with three temporal layers .....	54
Figure 5.3: Hierarchical coding with temporal id and disposable b frames combined .....	54
Figure 5.4: R-D performance for Crowd sequence .....	56
Figure 5.5: Prediction structure for a three layer temporal scheme using H.264 MVC. Straight lines indicate temporal prediction and dashed lines indicate inter-view prediction. ....	58

Figure 5.6: Default arrangement of reference picture list considering current slice as P5 for the enhancement layer prediction structure shown in Fig.5.5. It is assumed that the pictures P1 and P0 are coded as long term pictures. .58

Figure 5.7: Default state of reference picture list before the coding of picture P5 with respect to right view prediction structure shown in Fig.5.5. All reference pictures are coded as short-term pictures and three temporal layer scheme is employed.....59

Figure 5.8: Modified reference picture list before coding of P5 to suit temporal scalability requirement for the right-view prediction structure shown in Fig.5.5. Temporal level of P5 is TL1 and the reference pictures should be present at TL0 and TL1. P5' is the inter-view picture which is also coded at temporal level TL1.....59

Figure 5.9: Flow chart for the default list construction process for H.264 variants and the italicized text depicts the proposed modifications [P1].....60

Figure 6.1: Various prediction schemes employed by the SMVC method. Dotted lines represent temporal prediction, dashed lines represent inter-view prediction and curved dash-dot lines represent inter-layer prediction. ....65

Figure 6.2: PSNR and speed evaluation process between H.264 SVC and HEVC based up-sampling filters. Process shows only the left view of the crowd sequence and the same will be repeated for right view as well and the average between the two values will be computed.....69

Figure 6.3: Hierarchical prediction structure and the QP cascading scheme used in the simulation setup.....71

Figure 6.4: R-D performance curve for Flamenco sequence for the spatial scalability case.....73

Figure 6.5: R-D performance curve for Race sequence for the spatial scalability case .....74

Figure 6.6: R-D curves for Akko & Kayo sequence: Spatio-temporal scalability .....78

Figure 6.7: R-D curves for Crowd sequence: Spatio-temporal scalability .....78

Figure 6.8: R-D curves for Flamenco sequence: Spatio-temporal scalability .....79

Figure 6.9: R-D curves for Race sequence: Spatio-temporal scalability.....79

Figure 7.1: Hierarchical prediction structure and the QP cascading scheme used in the simulation setup.....83

Figure 7.2: R-D performance curve for Flamenco sequence for the coarse grain scalability case.....84

Figure 7.3: R-D performance curve for Race sequence for the coarse grain scalability case .....84

Figure 7.4: Packet arrangement in H.264 MVC.....86

Figure 7.5: Interpretation process for temporal identifier of base view .....87

Figure 7.6: Enhancement view at 1/2nd of base view frame rate.....87

Figure 7.7: Enhancement view at 1/4th of base view frame rate.....88

Figure 7.8: R-D curves for Akko & Kayo sequence: CGS-temporal scalability.....91

Figure 7.9: R-D curves for Crowd sequence: CGS-temporal scalability .....91

Figure 7.10: R-D curves for Flamenco sequence: CGS-temporal scalability .....92

Figure 7.11: R-D curves for Race sequence: CGS-temporal scalability .....92

Figure 8.1.a: FMO interleaving mode .....95

Figure 8.1.b: FMO Checker board mode.....95

Figure 8.2: Depth image based concealment scheme [78] .....96

Figure 8.3: Error concealment scheme using depth and color image [78] .....97

Figure 8.4: 4-D arrangement of MVV data [83] .....98

Figure 8.5: R-D curves for Akko & Kayo sequence: CGS error robustness .....102

Figure 8.6: R-D curves for Crowd sequence: CGS error robustness.....102

Figure 8.7: R-D curves for Flamenco sequence: CGS error robustness.....103

Figure 8.8: R-D curves for Race sequence: CGS error robustness.....103

Figure 8.9: CGS Subjective quality – Akko & Kayo: Originally encoded (PSNR -42.9 dB) on the left and concealed frame on the right (PSNR – 38.6 dB) .....	104
Figure 8.10: CGS Subjective quality – Crowd: Originally encoded (PSNR – 40.7 dB) on the left and concealed frame on the right (PSNR – 35.5 dB) .....	104
Figure 8.11: CGS Subjective quality – Flamenco: Originally encoded (PSNR – 42.8 dB) on the left and concealed frame on the right (PSNR – 37.8 dB) .....	104
Figure 8.12: CGS Subjective quality – Race: Originally encoded (PSNR – 41.58 dB) on the left and concealed frame on the right (PSNR -39.23 dB).....	105
Figure 8.13: R-D curves for Akko & Kayo sequence: SS error robustness .....	107
Figure 8.14: R-D curves for crowd sequence: SS error robustness .....	107
Figure 8.15: R-D curves for Flamenco sequence: SS error robustness .....	108
Figure 8.16: R-D curves for Race sequence: SS error robustness.....	108
Figure 8.17: SS Subjective quality – Akko & Kayo: Originally encoded frame (PSNR – 43.38 dB) on the left and concealed frame (PSNR – 37.6 dB) on the right .....	108
Figure 8.18: SS Subjective quality – Flamenco: Originally encoded frame (PSNR –43.5 dB) on the left and concealed frame (PSNR – 35.7 dB) on the right .....	109
Figure 8.19: SS Subjective quality – Crowd: Originally encoded frame (PSNR – 40.9 dB) on the left and concealed frame (PSNR – 32.25 dB) on the right .....	109
Figure 8.20: SS Subjective quality – Race: Originally encoded frame (PSNR – 42.4 dB) on the left and concealed frame (PSNR – 37.5 dB) on the right .....	109

## List of Tables

Table I: Bandwidth transmission requirements for the transmission of uncompressed video .....	16
Table II: Comparison of various schemes from the different SMVC requirement perspective .....	30
Table III: Coding tools used in JM for the simulation of temporal scalability.....	55
Table IV : Bjontegaard R-D performance numbers, DPSNR is in units of dB and DRate is in units of kbps.....	56
Table V: R-D performance comparison between existing and modified H.264 MVC schemes for a three temporal layer and stereoscopic video case.....	62
Table VI: PSNR and speed of conversion numbers for H.264 SVC and HEVC based filters. Performance measured on an Intel core i3 laptop clocked at 2.4 GHz with 4GB RAM .....	70
Table VII: R-D performance figures of simulcast and proposed SMVC methods for spatial scalability .....	72
Table VIII: R-D performance figures for base layer and base view of base layer for spatial scalability.....	73
Table IX: Compression ratio and encoder processing speed for Spatial Scalability .....	75
Table X: R-D performance results for Spatio-temporal scalability .....	77
Table XI: Compression ratio and encoder processing speed for Spatio-temporal scalability .....	80
Table XII: R-D performance figures of simulcast and proposed SMVC methods for coarse grain scalability .....	82
Table XIII: R-D performance figures for base layer and base view of base layer for coarse grain scalability .....	83
Table XIV: Compression ratio and encoder processing speed for coarse grain scalability .....	85
Table XV: Percentage bit rates and average frame extraction time for IBBP case. Frame rates are with respect to original encoded frame rate. $\frac{1}{2}$ means 50% with respect to original frame rate. ....	88
Table XVI: Percentage bit rates and average frame extraction time for IPPP case. Frame rates are with respect to original encoded frame rate. $\frac{1}{2}$ means 50% with respect to original frame rate. ....	89
Table XVII: R-D performance results for CGS-temporal scalability.....	90
Table XVIII: Compression ratio and encoder processing speed for CGS-temporal scalability.....	93
Table XIX: R-D performance results error robustness of CGS scalability .....	101
Table XX : R-D performance results error robustness of spatial scalability .....	106

## List of Acronyms

<b>Term</b>	<b>Expansion</b>
3DTV	Three Dimensional Television
4G	Fourth Generation
ADSL	Asymmetric Digital Subscriber Line
AVC	Advanced Video Coding
BBC	British Broadcasting Corporation
CABAC	Context Adaptive Binary Arithmetic Coding
CAVLC	Context Adaptive Variable Length Coding
CGS	Coarse Grain Scalability
CIF	Common Intermediate Format
CODEC	Coder decoder
DB	Decibels
DCT	Discrete Cosine Transform
DCVF	Disparity Compensated View Filtering
DMVE	Decoder Motion Vector Estimation
DPB	Decoded Picture Buffer
DPSNR	Delta Peak Signal to Noise Ratio
DV	Disparity Vector
DVB	Digital Video Broadcasting
DWT	Discrete Wavelet Transform
EDGE	Enhanced Data GSM Environment
ELXSI	Electronics System Integration
ESCOT	Embedded Sub band Coding with Optimal Truncation
FEC	Forward Error Correction
FIR	Finite Impulse Response
FMO	Flexible Macro block Ordering
FPA	Frame Packing Arrangement
FPS	Frames Per Second
FTV	Free view point TV
GoGoFs	Group of Group of Frames
GPRS	General Packet Radio Service
GSM	Global System for Mobile Communications
HD	High Definition
HEVC	High Efficiency Video Coding
HRD	Hypothetical Reference Decoder
HTC	High Tech Computer Corporation
IDR	Instantaneous Decoding Refresh
IEC	International Electro technical Commission
INTEL	Integrated Electronics
IP	Internet Protocol
IP core	Intellectual Property core
IPTV	Internet Protocol TV
IRD	Integrated Receiver Decoder
ISO	International Standards Organization
IT	Integer Transform
ITU	International Telecommunications Union

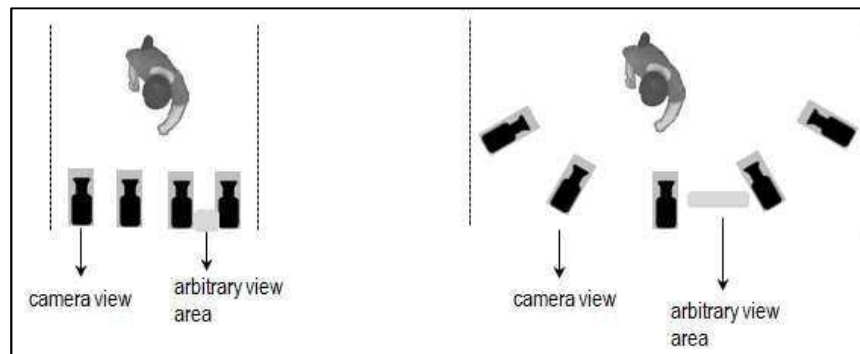
JM	Joint Model
JMVC	Joint Multi Video Coding
JSVM	Joint Scalable Video Model
JVT	Joint Video Team
LTE	Long Term Evolution
MANE	Media Aware Network Element
MB	Macro Block
MBAFF	Macro Block Adaptive Field Frame
MC	Motion Compensation
MCTF	Motion Compensated Temporal Filtering
ME	Motion Estimation
MHP	Multi view High Profile
MMCO	Memory Management Control Operations
MPEG	Moving Picture Experts Group
MV	Motion Vector
MVC	Multi view Video Coding
MV-HEVC	Multi view High Efficiency Video Coding
MVV	Multi View Video
NAL	Network Abstraction Layer
PAFF	Picture Adaptive Field Frame
POC	Picture Order Count
Power VR	Power Video Rendering
PPS	Picture Parameter Set
PSNR	Peak Signal to Noise Ratio
QCIF	Quarter Common Intermediate Format
QP	Quantization Parameter
RPLR	Reference Picture List Reordering
SDK	Software Development Kit
SEI	Supplemental Enhancement Information
SHP	Stereo High Profile
SMV	Super Multi View
SMVC	Scalable Multi View Video Coding
SPS	Sequence Parameter Set
SS	Spatial Scalability
STB	Set Top Box
SVC	Scalable Video Coding
TS	Temporal Scalability
TV	Television
UHD	Ultra High Definition
UMTS	Universal Mobile Telecommunications System
VCEG	Video Coding Experts Group
VCL	Video Coding Layer
VDX	Video Decoder Multi Standard
VGA	Video Graphics Array
VLC	Variable Length Coding
VLC	Variable Length Coding
VUI	Video Usability Information
VUI	Video Usability Information
XGA	Extended Graphics Array

# Chapter 1

## Introduction

We live in a multimedia world, where more than half of the internet bandwidth is used for video based applications. Deployment of video services have made a lot of progress in the recent times and as a result there is a renewed interest in the 3D video services as well. In the early days, 3D video was used majorly in the cinemas where the size of the screen was an added advantage in providing the “3D feel” to the audience. But in recent times, 3D video made entry into a number of consumer electronics applications such as 3D gaming consoles, 3D enabled mobile phones/tablets/laptops, 3D televisions, 3D movie play back devices (Eg: blu-ray players) etc. Hence there is considerable momentum has been setup in the deployment of 3D video services. In general, 3D video applications can be grouped under three broad categories:

- Three-dimensional TV (3DTV)
- Free view point TV (FTV)
- Immersive video conferencing



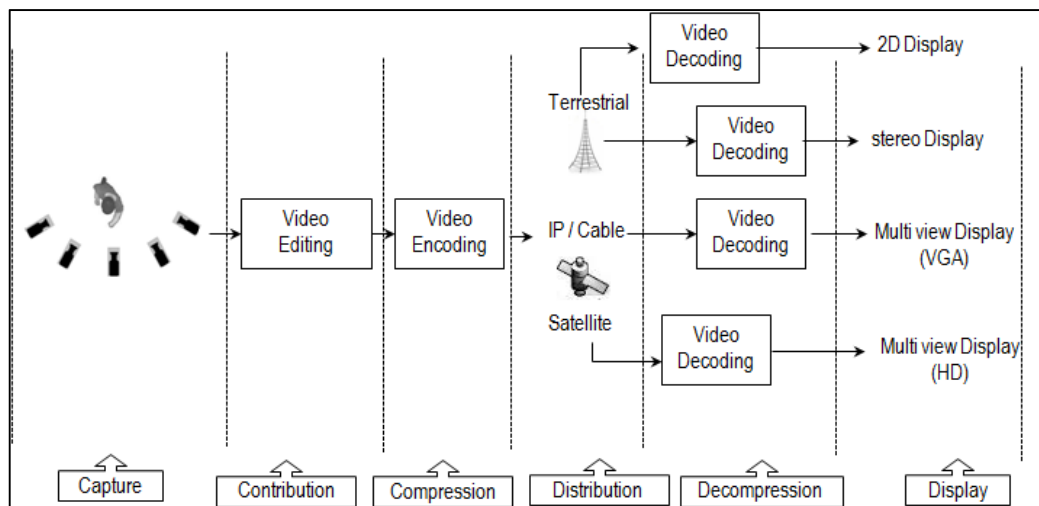
**Figure 1.1:** Camera arrangements for capturing sequences with multiple camera system. Camera setups are: 1-D line (left) and 1-D arc (right).

There are a number of camera arrangements exist to capture the video sequence in a multi-view system, among them a one-dimensional multi-camera arrangement is suitable for horizontal-parallax systems, where the head movement of the user is assumed to be in the horizontal axis only. As this assumption holds good for both 3DTV and 3D-video conferencing, the 1-D line and 1-D arc [1] based systems are widely used and these arrangements are depicted in Fig.1.1.

3DTV provides immersive experience to the user by transmitting videos captured using a large number of synchronized cameras arranged in 1-D line or 1-D arc arrangement as shown in Fig.1.1. In these systems, number of available viewpoints is equal to the number of cameras used to capture the scene. For example, with respect to

the 1-D line arrangement shown in Fig.1.1, the number of viewpoints is equal to four and for the 1-D arc arrangement, the number of viewpoints is equal to five.

FTV enable the users to view a 3D world by freely changing their viewpoints as if they were there, this is possible as, in theory an infinite number of viewpoints could be created in a FTV system [2 - 3]. In other words, the number of available viewpoints is greater than the number of cameras used to capture the scene and the "virtual" views need to be synthesized from the existing limited number of input cameras. In Fig.1.1, the virtual synthesized views are marked as "arbitrary view areas". Super Multi View (SMV) system could be considered as the super-set of FTV [4], where the number of virtual synthesized views will be very high when compared to the FTV. Thus, multi view video captured by a number of synchronized cameras, from different viewpoints, comprises rich information of a scene and is well poised to be used as a signal of information for new 'content - rich' visual media such as 3DTV, FTV and SMV TV. The overview of the multi view video system, which is used in SMV TV, FTV and 3DTV scenarios from capture to display, is shown in Fig.1.2.



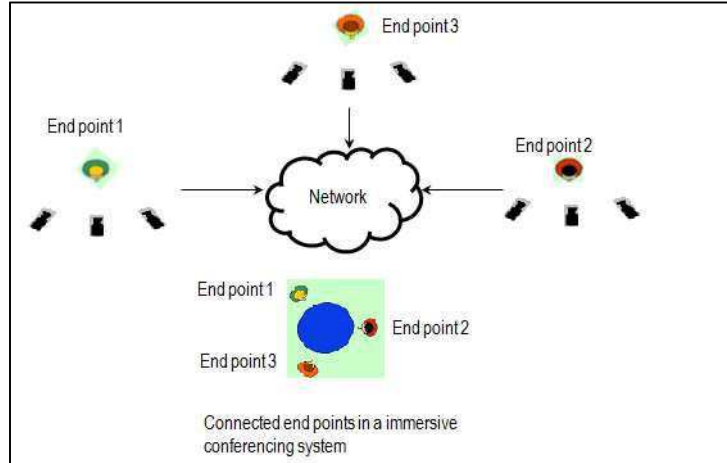
**Figure 1.2:** Overview of multi view video system

The scene is captured by a large number of synchronized cameras arranged in a particular fashion as mentioned in Fig.1.1. For the example shown in Fig.1.2, a 1-D arc based system is used and video frames will be sent to the studio for further editing, this part of the system is known as "contribution". High speed optical fiber cables will be used between the camera and studio, where keeping the quality of the frames will be the major focus at this stage. After editing, video data will be compressed and distributed to the user by using a variety of networks such as satellite, cable, Internet Protocol (IP) etc. and this part of the system is known as "distribution". At the user end, the data will be decoded and outputted based on the display system (2D / stereo 3D / multi view 3D) and this part of the system is known as "consumption". The major challenge in the multi-view video system is at the distribution stage as outlined below.



Availability of 3D application televisions are on the rise in the last few years as they not only enable the user to watch the content using set-top-boxes, but also allow the user to watch the contents such as 3D movies via internet. For example, Disney streams the 3D movies to internet-connected LG TVs using the “3D world” application which comes along with the LG TV [5]. Apart from the movie contents, the user could also watch the available 3D video contents from sources such as YouTube and Vimeo over internet. There are a number of options exist to deliver these 3D contents over internet, such as internet protocol TV (IPTV), over the top streaming (OTT) and video on demand (VOD), which differs from each other in terms of available bandwidth. We live in an age where TV shows are watched not only in televisions but also in smart phones [6], tablets, note books and laptops, where screen sizes are different from each other. The explosive growth in the internet usage is inarguably one of the most important milestones of the last decade and this made “connected anywhere, everywhere” a reality, even on the move, thanks to the evolution of wireless technology. As a result the 3D video capable end devices such as networked 3DTVs, smart phones, tablet PCs, Note books, and Laptops are connected over a wide variety of IP network connectivity options such as GPRS, EDGE, LTE, WI-Max, UMTS, ADSL, 4G and this makes distribution of the content a challenging task. In summary, the multi view video content needs to be distributed to a number of devices with various display and processing capabilities, connected over a number of heterogeneous networks with various bandwidth capabilities.

Another use case of the multi-view system is the immersive 3-D video conferencing system [7], as shown in Fig.1.3. The conferencing participants are located in various geographies and connected with each other using the video conferencing system. Multiple cameras are mounted in every location and the content will be compressed before sent to other participants. At the same time content received from the other users will be decompressed and displayed, providing the immersive feel to the participants. Similar to the case of TVs, 3D-capable mobile phones and tablets are used in the 3-D video conferencing system and the notable difference is that the latency of the system should be less in the case of immersive 3-D video conferencing system compared to the 3DTV. In summary, even in the case of immersive 3-D video conferencing, a number of devices with various display and processing capabilities needs to be connected over a number of heterogeneous networks with various bandwidth capabilities and interoperability is a challenging task in this case too.



**Figure 1.3:** Immersive 3D conferencing system

From the above two examples, it is clear that there is a need for a scalable multi-view video scheme, which satisfies the following requirements.

**Compression efficiency:**

A multi view system employs a number of cameras to capture the scene and Fujii.et.al in [8] used up to 100 cameras to capture the scene in VGA (640x480) resolution. The bandwidth requirements to transmit the uncompressed videos for different resolutions, frame rates w.r.t 4:2:2 color space is depicted in Table I. It could be seen that a resolution as low as VGA at 30 frames per second needs a bandwidth of 15 Gbps to transmit the uncompressed video, whereas the maximum available fixed-line bandwidth as of today are in hundreds of Mbps. On the other hand, the maximum available wireless bandwidth is around 50Mbps. This requires the scalable multi view video coding scheme to provide superior compression efficiency to accommodate a number of users in the given internet spectrum.

**Table I:** Bandwidth transmission requirements for the transmission of uncompressed video

Picture width (in Pixels)	Picture height (in Pixels)	Number of cameras	frame rate (fps)	Bandwidth (Gbps)
640	480	100	30	15
640	480	100	60	29
1280	720	100	30	44
1280	720	100	60	88
1920	1080	100	30	100
1920	1080	100	60	199
2048	1556	100	30	153
2048	1556	100	60	306
3840	2160	100	60	796

**Backward compatibility:**

The scalable multi view video scheme should provide backward compatibility such that the legacy 2D and existing 3D devices could be served using this scheme. In other words, in a setup which contain a number of legacy 2D and/or 3D devices, the scalable multi view video compatible devices should be seamlessly introduced and all the devices could be connected to each other using a single encoded bitstream.

**Re-use of existing infrastructures:**

The deployment cost for the infrastructures is high and maximum re-use of existing infrastructure will be very important for any new compression scheme. This will allow the service providers to introduce a new compression scheme with limited or no additional cost to the end user, which helps in early adoption of the technology.

**Stereo 3D support:**

Stereo 3D refers to the scenario where there are two views captured by right and left cameras which mimic the human eye. Stereo 3D is an important use case, which is considered as a starting point for the deployment of multi view video. Thus, the scalable multi-view video coding scheme should consider the support for stereo 3D compression and decompression as a special use case.

**Ease of packet extraction:**

The packet extraction time at the media aware network element (MANE) should be less as deep inspection of the packets will consume computing resources at the server in addition to the increased latency from capture to display. As noted in the case of immersive of 3D videoconferencing system, capture to display latency should be very minimal and to achieve this, the packet extraction process at the server should be as much light-weight as possible. In theory, the information to classify the packet should be present in the packet header, such that a small parser at MANE will perform the packet extraction process.

**Scalability support:**

As noted in both the multi-view TV and immersive video conferencing based applications, there are a number of devices with various display resolution/ frame rate/ bit rate / processing capabilities needs to be connected with each other. Thus the scalable multi view coding scheme should provide a number of scalability options such as view scalability, temporal scalability, bit rate scalability and spatial scalability. The various scalability requirements ensure that the proposed scheme is capable of accommodating a number of devices with different capabilities and also enable them to talk to each other in a seamless fashion.

**Single loop decoding:**

In a layered video coding architecture, multiple loop decoding refers to the scenario where the decoding process involves decoding of more than one layer, including the target layer. Multiple-loop decoding is one of the major reasons for the failure of previous standards such as MPEG-2 multi view profile [9]. In single loop decoding, to decode a target layer that depends on a number of lower layers, only the target layer needs to be fully decoded, while for the reference (lower) layers only parsing and minimal decoding are needed [10]. The single loop decoding process not only improves the decoder speed, but it also helps in reducing the memory requirements of the decoder. Hence, the scalable multi view coding scheme should support single loop decoding, with as minimal decoding of non-target layers as possible.

**Error robustness:**

The scalable multi view coding scheme should aid the decoder in error-prone scenarios. For example, when the video packets belong to a particular layer are lost, the resultant error could be minimized by using other layer's packets in addition to the other intra-layer packets. It could be noted that for this error robustness process also, the scalable multi view coding scheme should ensure that multiple loop decoding is not required.

**Encoder-decoder mismatch:**

The scalable multi view coding scheme should ensure that there is no drift introduced between the encoding and decoding schemes. The encoder-decoder drift results in poor video quality, which could only be alleviated by sending an intra-refresh picture at regular intervals or by using mechanisms such as gradual decoding refresh, where parts of the frame are refreshed periodically. It could be noted that both the schemes requires additional transmission bandwidth.

**Inter-operability:**

The scalable multi view coding scheme should specify the bit-stream syntax, such that any scalable multi view video coding compliant device can be connected to an another scalable multi view video coding compliant device in a seamless fashion, allowing the devices to inter-op with each other.

**Objectives and outline of the thesis:**

The objective of this thesis is to introduce a scalable multi view video coding scheme which satisfies the above requirements. In short, the compression part of the multi view video system along with the signaling mechanisms is dealt in the proposed scalable multi view coding scheme and other areas of the system such as view synthesis, rendering of the decoded data are considered as out of scope. The rest of thesis is organized as follows, Chapter.2 provides the background literature in scalable multi view video coding, Chapter.3 provides an overview of H.264

Multi view video coding, Chapter.4 introduces the proposed scalable multi view coding scheme, Chapter.5 depicts the temporal scalability provided by the scalable multi view coding scheme, Chapter.6 illustrates spatial scalability and combined spatio-temporal scalability, Chapter.7 presents the bit-rate (both coarse grain and medium grain bit rate) scalability and combined coarse grain-temporal scalability, Chapter.8 explains the error robustness provided by the scalable multi view coding specification and Chapter.9 presents the conclusions and future scope of work.

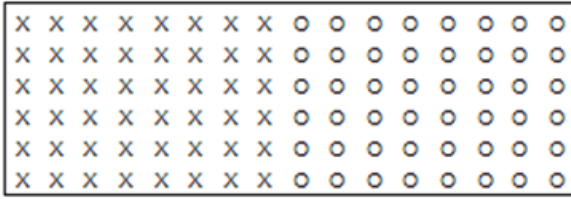
## **Chapter 2**

### **Back ground**

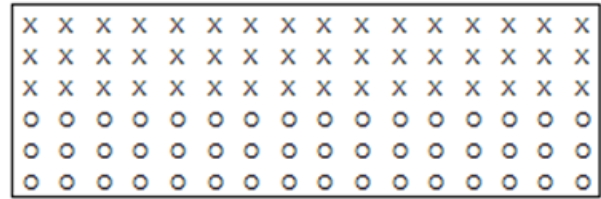
This chapter provides a brief introduction to the existing multi-view video compression schemes and compares them from a number of different perspectives such as compression efficiency, backward compatibility, stereoscopic 3D support, re-use of existing infrastructure, single loop decoding support and support for various scalability options. The primary objective is to find the best multi-view video compression scheme among the available and enhance the same to meet the various requirements mentioned in Chapter.1. The secondary objective is to enhance the frame-compatible services (spatial interleaving methods) as these methods will be used in the first phase of 3DTV deployments. The spatial and temporal interleaving based methods are discussed first, followed by the wavelet based scheme. Finally, hybrid coding based multi-view coding scheme is provided along with the table which depicts the summary of all the schemes from the different perspectives as noted above. It could be noted that in all these comparisons, H.264 video compression scheme is considered as the base as it is deployed widely and the newly introduced high efficiency video coding (HEVC) compression is considered wherever applicable. This chapter summarizes the publications [J1] and [C1].

#### **2.1. Spatial interleaving using Frame Packing Arrangement**

Spatial interleaving refers to the scenario where the information from different cameras (view points) will be resized and packed in to a single frame. The major advantage of spatially interleaving the left and right view video frames to form a single 2D frame is that it allows the existing contribution and distribution infrastructures to be used effectively for the seamless introduction of stereoscopic 3D video services. Though there are a number of spatial interleaving frame-compatible formats, namely Side-by-Side (SbS), Top-And-Bottom (TaB), Row interleaved, Column interleaved and Checkerboard [11] exist, only SbS and TaB methods are used widely and these are illustrated in Fig.2.1.a and Fig.2.1.b respectively. It could be noted that the compression performance also depends on the efficiency of the up/down sampling filters used to form the 2D frame in addition to the compression scheme used and additionally the ‘two frames packed in a single frame’ information needs to be conveyed to the decoder for the appropriate display.



**Figure 2.1.a:** SbS frame compatible format. ‘X’ denotes the samples from left view and ‘O’ denotes the samples from right view.



**Figure 2.1.b:** TaB frame compatible format. ‘X’ denotes the samples from left view and ‘O’ denotes the samples from right view.

The Frame Packing Arrangement (FPA) Supplemental Enhancement Information (SEI) was introduced in H.264/AVC as a method to signal the ‘two frames packed in a single frame’ information and has been strongly embraced by industry as the preferred method for deployment in broadcast stereoscopic 3D / mobile 3D video services [12]. DVB specification suggests that 30/25 Hz frame compatible Plano-stereoscopic 3DTV services will be provided by either SbS/TaB approach [13]. BBC uses the SbS approach to squeeze the left and right view pictures to form the 3D video [14] and HTC Evo also uses the SbS method to form the stereoscopic 3D frame [15]. There are several problems associated with FPA SEI messages in H.264/AVC; Since the FPA SEI messages are introduced in an amendment which is released after the introduction of H.264/AVC specification, there were 2D decoders already exist in the market. Another problem is that even though the 2D decoders were implemented according to a version of H.264/AVC which has the support for FPA SEI, decoding of SEI messages are not mandatory as part of the standard and as a result even the standard-compliant 2D decoders will not decode the FPA SEI and output the decoded pictures directly as the packed frame. Hence, though the H.264/AVC SEI FPA facilitates the introduction of stereo 3D using the existing infrastructure, the interoperability issues still exist. Some broadcasters chose the ‘simulcast’ of base view (for 2D receivers) and FPA SEI based 3D video, to resolve these issues at the expense of increased spectrum usage. In the UK, BBC introduced 3D broadcasting using the ‘spare’ available spectrum on BBC HD. The left and right view frames are arranged in SbS method to form the 3D frame and is available on BBC HD [14] and in order to solve the backward compatibility issues, 2D version of the same programs are aired on ‘BBC One HD’.

Situation is better with HEVC, as the standard included the support for FPA SEI messages in the draft version itself. Several contributions have been made to resolve the issues that arose in H.264/AVC. One proposal is to make the decoding of the FPA SEI process mandatory by adding an additional flag in the Video Usability Information (VUI) [16]. When the flag is set, all decoders shall decode SEI as a mandatory process, so as to produce identical output even when the views are packed together as a single frame. Another proposal is to use the cropping information in Sequence Parameter Set (SPS) along with the flag in VUI. For the 2D decoders,

which support the FPA SEI, when the FPA flag is set, the cropping syntax (which specifies the offset to crop the base view) along with the `pic_cropping_flag` is used to extract the base view. The 2D decoders, which do not support FPA SEI, also will extract the base view as a result of the cropping information in SPS. Introduction of an additional flag (`general_non_packed_only_flag`) in profile, tier and level syntax [17] will also ensure the same. When FPA is used, this flag will be reset, which indicates to the decoder that two views are squeezed in to one single frame and the flag will be set for 2D case. When the two views are packed in to a single frame, the amount of information (such as number of max luma samples) makes the level of the bitstream to be set relatively high. This will prevent the 2D decoders in low-end devices from decoding the stream, even though the relevant information (base view) is within its capable level limits. One possible approach is to introduce a SEI message called Independent decodable region SEI [18], which will contain the cropping information, and optionally the ‘new’ level information, which correspond to the base view. This will expand the possibility that decoders with low-“level capability” could also decode the stream.

In summary, though the spatial interleaving method enables the introduction of stereo 3D using the existing infrastructures, there is no support for scalability options and compression performance is inferior when compared to the other methods [J1, C1]. Also, there are a number of interoperability issues exist and more than two views could not be supported. In order to improve the coding efficiency and parallel processing of data with the HEVC coding specification, the frames from left and right cameras are encapsulated in two different tiles with filters disabled across the tile boundary. The intuition behind this is that due to independent nature of the data, the loop filters will only introduce noise instead of smoothening the pixels present across the tile border. On the other hand, if the in-loop filters are turned off, the complete decoding and encoding of frames could happen in parallel without affecting the rate-distortion performance. Experimental results show that the proposed method of encapsulating the two views into two different tiles with filters turned off across the tile boundaries improves the decoding performance by almost 50% with approximately the same rate–distortion performance when compared to ‘no-tiles’ case [J1].

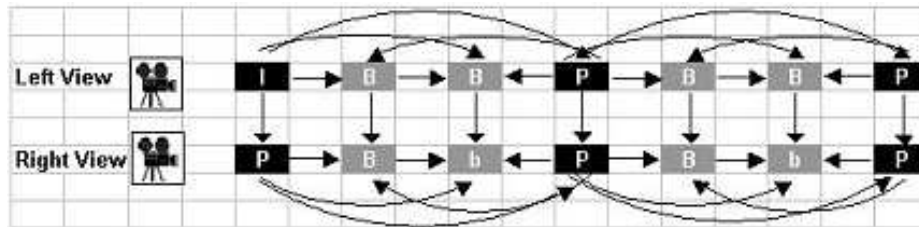
## **2.2. Temporal interleaving using SVC**

Temporal interleaving refers to the scenario where frames from different cameras (view points) will be encoded alternatively, but coded at different temporal layers. For example, consider the stereoscopic case, where the left view and right view frames will be encoded sequentially, with left view frame placed in the lowest temporal layer and right view frame in the enhancement layer. H.264 SVC compression scheme is adopted by DVB specification and is deployed in many STB / IRDs. SVC follows a layered approach, where the base layer is AVC compatible and the enhancement layers will be used to serve different SVC end points. H.264 SVC supports spatial, quality and temporal scalabilities for 2D video and it does not have the support for view scalability. The view scalability,

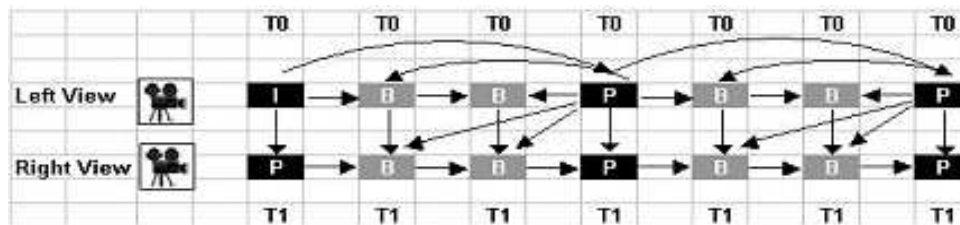


from the perspective of stereo 3D, indicates that base layer contains the left view pictures and the enhancement layer contains the right view pictures. An example prediction structure for view scalability is shown in Fig.2.2. The same could be realized by using H.264 SVC using the temporal scalability (SVC-TS) with left view pictures encapsulated in the lowest temporal layer and enhancement layer contain the right view pictures as shown in Fig.2.3.

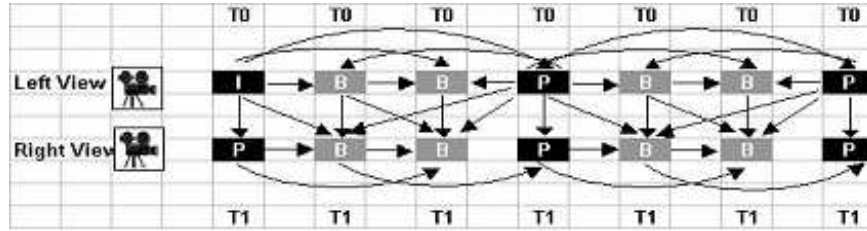
From the compression perspective, for the prediction structure illustrated in Fig.2.2, there will be loss in compression efficiency when SVC is used (as illustrated in Fig.2.3) due to the availability of less number of “temporal” reference pictures and this could be compensated to some extent by having more number of reference pictures per reference picture list. Fig.2.4 depicts the SVC-TS prediction structure when ‘4 reference frames’ are used per reference picture list. It could be seen that there are more ‘temporal predictors’ available when compared to prediction structure in Fig.2.3. SVC also has coding tools like residual prediction, MV prediction etc. at MB level, however for ‘temporal scalability only’ case, these tools are not required and could be tuned ‘off’ using fields in the packet header. Thus it could be expected that SVC-TS based approaches with sufficient number of reference frames would produce R-D performance that is comparable to existing state-of-the-art multi view specifications.



**Figure 2.2:** Prediction structure with 2 reference frames / list for view scalability



**Figure 2.3:** SVC- TS prediction structure with 2 reference frames / list



**Figure 2.4:** SVC-TS prediction structure with 4 reference frames / list

From the perspective of ‘resolving backward compatibility issue’, though temporal interleaving cannot resolve fully, it could be addressed to a certain extent by using the number of STBs available in the market with H.264 SVC support. Based on the capability of the TV, STB can either produce the 3D view or can output the base 2D view by extracting only the base temporal layer frames. But clearly the existing 2D infrastructure needs changing in order to accommodate this method. HEVC specification included the support for temporal scalability in the draft specification and there are proposals [16], which resolve the problem of ‘back ward compatibility’ using a flag (`max_temporal_id_minus1`) in SPS along with an additional flag in VUI. The `max_temporal_id_minus1` flag will be set to `temporal_id` of the base view and the 2D decoders will output the base view, where-as the 3D decoders will output both the views using the combination of the above-mentioned flags (SPS and VUI flags).

In summary, the temporal interleaving method provides better compression method when compared to spatial interleaving method [C1], but is still inferior when compared to existing state-of-the-art methods [J1]. Also, only a limited number of views could be supported and there will be reduction in compression efficiency when the number of temporal layers (views) is increased to more than two.

### 2.3. Wavelet based methods

It is widely believed that wavelet based scalable video compression will offer better coding efficiency than hybrid coding based scalable video schemes. Wavelet based CODECS which could offer spatial and SNR scalabilities for multi view video have been proposed in the literature so far [19 - 21]. These methods usually apply motion compensated temporal filtering (MCTF) along the temporal axis first followed by disparity compensated view filtering (DCVF) along the view axis to de-correlate the data in the temporal-view axis before applying discrete wavelet transform (DWT) for spatial de-correlation. For example, Garbas and Kaup in [22] have applied the MCTF in 3 stages followed by the DCVF stage. Both MCTF and DCVF stage uses 5/3 orthogonal filters. Finally a 9/7 filter has been used to spatially de-correlate the data and the entropy coding of the wavelet coefficients has been done by ESCOT (Embedded Sub band Coding with Optimal Truncation) [23]. Due to the inherent property of the wavelets, scalability could be easily achieved. But these CODECS have been designed in such a way that the coding and decoding will be performed in independent groups of groups of frames (GoGoFs).

For example, Garbas and Kaup in [22] presented the experimental results for the GoGoFs size of 16 (4 temporal levels across 4 views). This will introduce a massive requirement in memory and increase system latency.

In order to achieve a spatially down-scaled version of the video in the 3-Dimensional wavelet based single view CODEC ( $1V \times (T+2D)$ ), the sub bands corresponding to the high pass (details) will be discarded. This will introduce the "drift problem (operational mismatch)". The major reason for this problem is that, at the encoder side the high pass components of a reference frame will partially predict the low pass components of the anchor frame. When these high pass details are discarded at the decoder side (to achieve spatial scalability), drift will be introduced. Other reasons will be the motion vector scaling, overlapped block motion compensation (OBMC), fractional pixel shifts [23]. These drifts will appear more prominently at the high bit rates and imposes an upper limit on the bit rate for the acceptable quality.

The same problem exists in the 4-Dimensional wavelet based multi view CODEC ( $N \times (T+V+2D)$ ). Here the individual resolutions are created by using spatial transforms/filters similar to hybrid coding based SVC. But there is a burden on compression efficiency because the redundancy between the resolutions will be high. Similar to hybrid coding based SVC, the lower resolution frames are used to predict the high resolution frames. The drawback is that in-band low pass prediction will be less efficient when compared to the inter-layer prediction in image domain, as only part of the low resolution frame is used for the prediction in the case of wavelet based scheme. Xiong et.al in [24] introduced a weighing factor to the low resolution frames in such way that the prediction efficiency will be improved. But the quality of the predictor may vary from frame to frame or across the sub-bands, hence the estimation needs to be done periodically and also it will cost considerable bits for the communication to the decoder. Garbas and Kaup in [22] has shown that the decoded picture quality could be improved when the quantization error is fed back in to the prediction loop such that the predictor is nothing but the reconstructed version of the lower layer frames weighed by an appropriate factor. But this violates one of the major requirements for the efficient scalable scheme, which is the single - loop decoding [25] for the reduced complexity and system latency. Since in this scheme fully reconstructed lower layer frame has been used as the predictor for the sub bands of the higher layer frames, multiple loop decoding cannot be avoided. Moreover, the wavelet based methods will not be backward compatible with the existing devices which uses legacy compression specifications such as H.264.

It has also been shown that with wavelet based CODECs, the memory requirements will increase dramatically and system latency will also increase and added decoder complexity. For example in MCTF CODECs, the update steps in the decoder are extremely complex, second only to the de-blocking filter operations. Markus Flierl et.al in [1], presented that the wavelet based 3D video compression is far below in maturity level when compared to hybrid coding schemes.

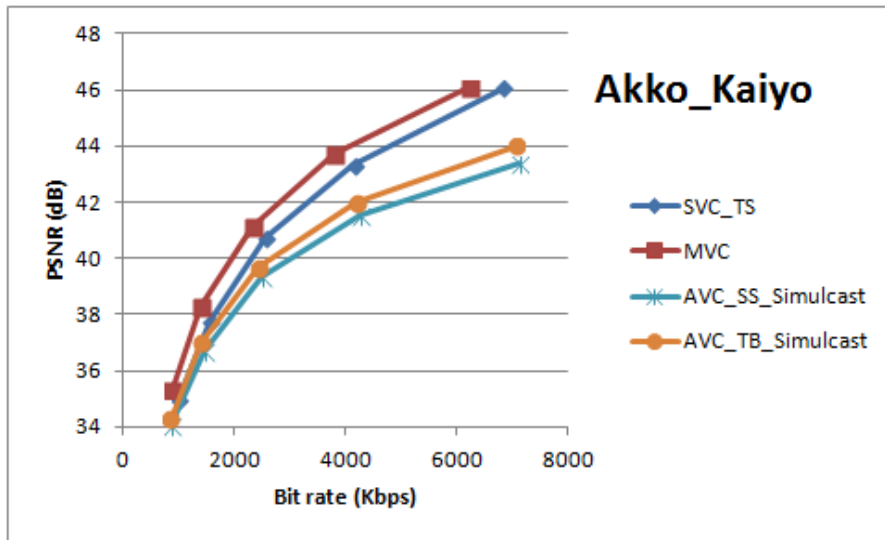
In summary, the wavelet based methods will not be backward compatible with the existing legacy devices as these devices are based on standards based compression scheme, there will be encoder-decoder mismatches and requires a complete change in the existing infrastructure. Also under certain conditions, multiple loop decoding will be required and has a massive memory and system latency requirements.

## **2.4. Hybrid coding based scalable 3D**

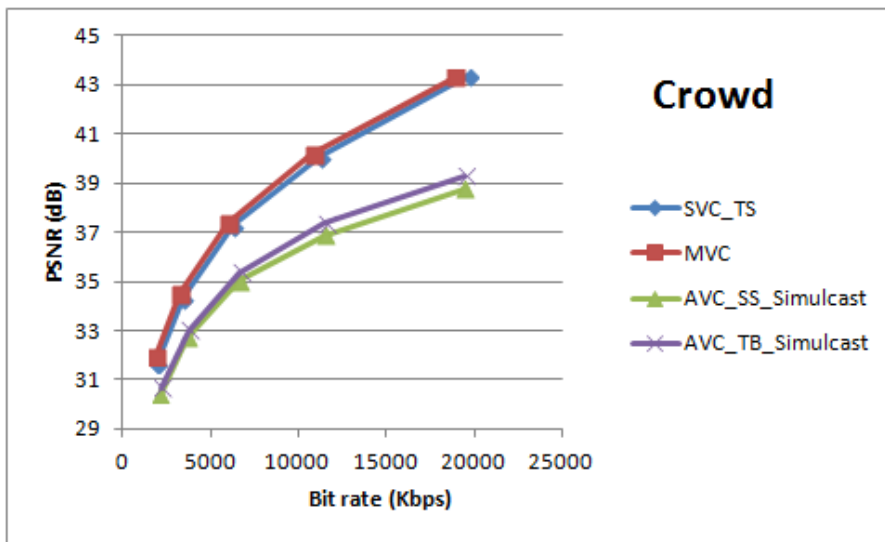
In hybrid coding based scalable compression scheme for 2D video, a single encoded bitstream will be used to serve a number of users with different display resolutions (spatial scalability), different frame rate requirements (temporal scalability) and with different bit rates (SNR scalability). Hybrid coding based scalable compression, at least for 2D video is not a new technology; in fact it is over two decades old. Many early standards such as MPEG-2/H.262 [26], H.263 [27] and MPEG-4 visual [28] have included tools to provide several important scalabilities, though these profiles have rarely been used [29]. The major reason for the failure of these profiles is that the decoding complexity was fairly high and these profiles usually come along with the loss in coding efficiency. Schwartz et.al in [30] has shown that the H.264 scalable video compression [31] has been designed in such a way that it overcomes all the shortcomings of the early standards except for few cases (E.g.: at synchronization points), it follows single loop decoding and the compression performance is comparable with that of single layer coding.

On the other hand, 3D video compression using hybrid video coding has been an active research topic for the last couple of decades at least. Early standards such as MPEG-2 already support compression of multi-dimensional video to some extent. To be specific, it supports only stereoscopic video compression in its Multi view profile [26]. Ohm in [32] has shown that the MPEG-2 standard could be used for providing the temporal scalability support, by coding the left eye information as the base layer and the right eye information in the enhancement layer. MPEG-4 also has the support for various multiple auxiliary components (MAC). But again due to the inadequate system support (including acquisition and display) and the limited compression efficiency, these multi view profiles have never been adopted by the industry. In order to overcome the afore-mentioned shortcomings and to meet other requirements specified in the requirements on multi-view coding [33], MPEG issued its “call for proposal” (cfp) [34] and started the evaluation process. This multi view video coding standard has been finalized in 2008 as an annexure to the H264 standard [35], which have been designed in such a way that it is backward compatible with H.264/AVC (base view). In other words, the base view supports 2D video compression and could be decoded by a H.264 AVC compliant decoder and the enhancement layer data contains multi view video. Chen et.al in [36], emphasized the fact that the MVC standard provides support for temporal and view scalability but does not support either spatial scalability or bitrate scalability. Thus it cannot be used to support multiple clients with different multi view screen resolutions over a heterogeneous network.

Rate-Distortion curves correspond to various schemes discussed above are depicted in Fig.2.5 to Fig.2.8 for both Akko & Kayo, crowd sequences [J1]. It could be seen that the hybrid coding based method (depicted as MVC in the graphs) provides superior compression efficiency when compared to other schemes such as spatial interleaving method (denoted as AVC\_SS\_Simulcast, HEVC\_SS\_Simulcast and AVC\_TB\_Simulcast ,HEVC\_TB\_Simulcast in the graphs), temporal interleaving method (marked as SVC\_TS in the graphs). In summary, the hybrid coding based method provides superior compression efficiency, encoder-decoder match, interoperability, reuse of existing infrastructure and supports single loop decoding. But it does not support different scalability options such as spatial, bit rate scalabilities. Table II summarizes the different SMVC requirements mentioned in Chapter.1 and the support provided by these methods (spatial interleaving, temporal interleaving and hybrid coding) for every requirement.



**Figure 2.5:** R-D curves for various H.264 based methods for akko & kayo sequence



**Figure 2.6:** R-D curves for various H.264 based methods for crowd sequence

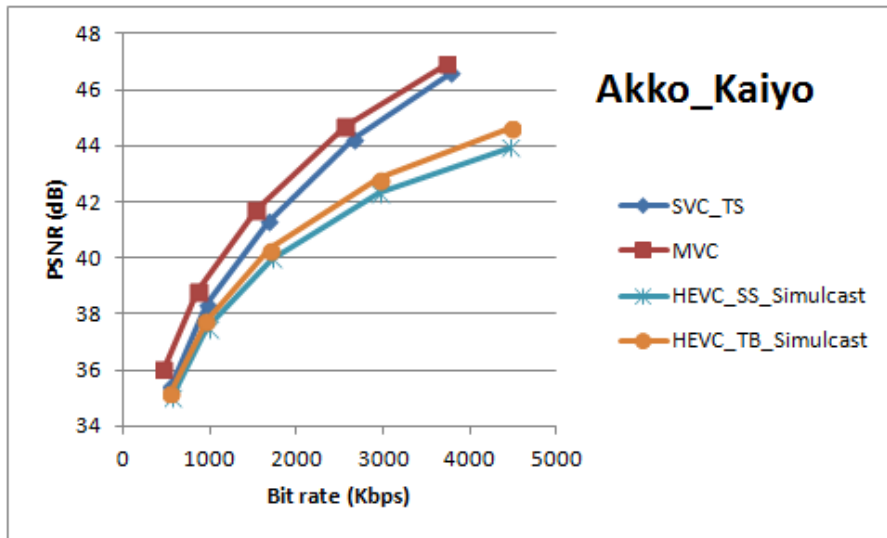


Figure 2.7: R-D curves for various HEVC based methods for akko & kayo sequence

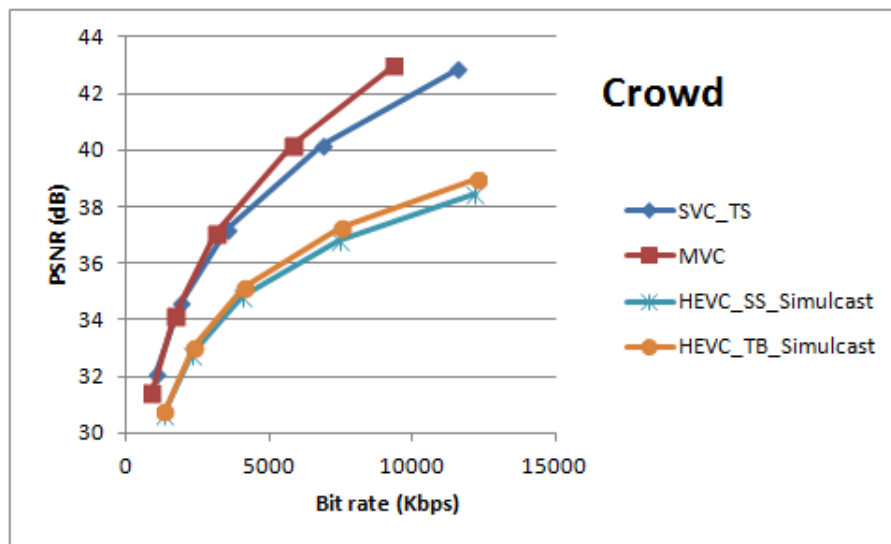


Figure 2.8: R-D curves for various HEVC based methods for crowd sequence

**Table II:** Comparison of various schemes from the different SMVC requirement perspective

SMVC requirement	Spatial interleaving scheme	Temporal interleaving scheme	Wavelet based scheme	Hybrid coding based scheme
Superior compression efficiency	×	×	×	√
Backward compatibility	×	×	×	√
Inter-operability support	×	×	×	√
Encoder-decoder match	√	√	×	√
Scalability support	×	×	√	X
Reuse of existing 2D infrastructure	√	√	×	X
Single loop decoding support	×	×	×	√

From the R-D curves, it is clear that the hybrid coding based method provides superior compression efficiency and Table II also depicts that all the SMVC requirements could be met except the scalability support and reuse of existing 2D infrastructure. In summary, a hybrid coding based fully scalable multi view scheme, which could offer all types of scalabilities will be the best solution to meet the various requirements mentioned in Chapter.1 for multi view video. But unfortunately very limited work has been done on this area so far. Anthony vetro et.al in [37] introduced a basic spatial scalable scheme and showed that it could not only alleviate the problem of decoding till the maximum resolution (beyond its display capability) and down size the video, but also provides a major boost to the robustness and flexibility of a heterogeneously networked system. In this thesis, a novel fully scalable multi view scheme using H.264 multi view video coding has been proposed to address all the SMVC requirements mentioned in Chapter.1. Though HEVC based multi view video compression has been standardized recently, it is not yet adopted by the industry. In other words, H.264 multi view video coding is still considered as the state-of-the-art compression for multi view video, hence the same is considered as the base for the thesis.



## **Chapter 3**

### **Overview of H.264 Multi view video coding specification**

This chapter provides a brief introduction to H.264 multi view video coding specification. As H.264 MVC is based on H.264/AVC, H.264/AVC specification is discussed first, followed by H.264 MVC standard. The intent of the chapter is to describe the features of H.264 MVC that are used in the coding tools of the proposed scalable multi view video coding specification.

#### **3.1. A brief overview of the H.264/AVC standard**

H.264/AVC is one of the specifications in the series of international video coding standards, jointly developed by ITU-T's Video Coding Experts Group (VCEG) and ISO/IEC's Moving Picture Experts Group (MPEG). The goal of the standard is to achieve the same quality video as that of MPEG-2, but by consuming half of the bitrate. There are a number of versions of the H.264/AVC standard, each adding new features to the specification. Some of the important versions include the following; Version 1.0 [38] refers to the first approved version of the standard containing Baseline, Main, and Extended profiles. Version 3.0 [39] refers to the addition of "Fidelity range extensions" amendment containing High, High 10, High 4:2:2, High 4:4:4 profiles. Version 8.0 [31] refers to the addition of "Scalable Video Coding" amendment, containing Scalable Baseline, Scalable High, Scalable High Intra profiles. Version 11.0 [35] refers to the addition of "Multiview Video Coding" amendment, containing Multiview High, Stereo High profiles. Version 21.0 [40] refers to the addition of "Multi view Plus Depth Coding" amendment, containing enhanced Multiview Depth High profile. The reference software for H.264/AVC (termed as the Joint Model (JM)) is based on C-language and is maintained by both ITU-T [41] and ISO/IEC [42]. The subversion (SVN) based software repository is maintained by HHI University [43].

The standard specifies bitstream syntax, semantics, decoding process but does not specify the encoding process. Since the decoding process is specified in the standard, all the encoders are expected to produce compliant bitstreams. The video data is encoded into a number of network abstraction layer units, which contain header and data parts. The header part contains information about the importance of the video data which is present in the succeeding NAL packet and will be used by the network element in scenarios such as bitstream adaptation. The NAL data could be classified as Video Coding Layer (or) non video coding layer data, where the non-video coding layer data may include information such as video sequence related data (sequence parameter set) (or) picture related data (picture parameter set). The advantage of segregating the video data from parameter sets is that these information could be sent during the service initialization time using secure channel, which also contributes to bitrate savings as the redundancy in sending the header information is removed.

The base unit of the H.264/AVC specification is the picture and the same could be encoded as a single frame (or) top and bottom fields. Frame coding is termed as progressive encoding and field coding is referred to as interlaced encoding, which is useful in high motion scenarios. The frame/field data contains both the luminance pixels and corresponding chrominance pixels; in case of 4:4:4 video format, the amount of chrominance pixels is equal to the luminance pixels. In case of 4:2:2 video format, the number of chrominance pixels is equal to the luminance pixels in the vertical direction but subsampled by a factor of two in the horizontal direction. Where as in the case of 4:2:0 video format, the amount of chrominance pixels is subsampled in both horizontal and vertical directions. In H.264/AVC, the frame/field data is divided in to a number of rectangular areas termed as slices. Slices are independent units with no sharing of data across them and they are further subdivided in to a number of 16x16 elements called Macroblocks. For example, in the case of 4:2:0 video format, a 16x16 macro block contains data corresponding to 16x16 luminance pixels, 8x8 blue channel chrominance pixels and 8x8 red channel chrominance pixels.

### **3.1.1. Profile and Level**

Profile defines a group of coding tools from the list of tools supported by H.264/AVC. For example, one of the coding tool supported by H.264/AVC is the Bi-directional frames (B-frames). Though B-frames improve the compression efficiency, it also increases system latency as the frames need to be stored before display. Thus the usage of B-frames will not be suitable for low latency applications and the low latency encoder can choose a profile which does not support the encoding of B-frames. On the other hand, level adds more details and imposes restrictions on the selected coding tool. For example, level specifies the number of macro blocks that could be encoded per second, which indirectly imposes the restriction on the maximum size of the encoded picture width and height. The profile and level information is encoded in the sequence parameter set and as mentioned earlier, the parameter sets are sent to the decoder during the session initialization time. This has several advantages, first the redundancy in encoding the same information in every picture header is removed, and second the media aware network element can match the decoder's capability with the parameter sets and determine whether the receiver is capable of decoding the compressed video content before the transmission of the packets and finally the decoder could reserve the required video memory from the media framework's memory pool well in advance.

Though H.264/AVC supports a number of profiles, only the high profile and constrained baseline profile are briefly mentioned below as these profiles are used in the H.264 MVC specification.

- Constrained baseline profile is suitable for applications where the system latency will be minimal, for example video conferencing and mobile applications. It supports encoding of Intra (I), Predictive (P) frames and all the other fundamental coding tools such as 4x4 integer transform, multiple reference

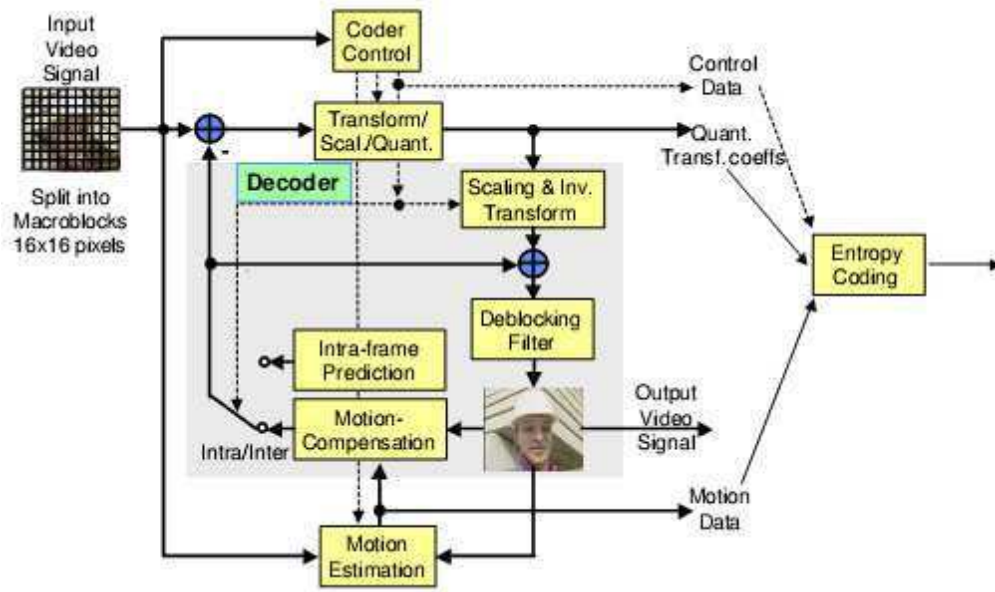
pictures etc. are supported. Since the latency is of prime concern, there is no support for the encoding of B-frames. Also, coding tools such as picture adaptive frame/field (PAFF) coding, MB adaptive frame/field (MBAFF) coding are not supported as the motion content of the objects are expected to be on the lower side. As a consequence, only progressive videos are considered as video source for encoding and other error resilient tools such as redundant slices are not supported. The compute intensive tools such as context adaptive binary arithmetic coding (CABAC) are also not supported, instead the "light weight" context adaptive variable length coding (CAVLC) is supported. In summary, this profile places more emphasis on latency compared to compression efficiency, error resilience.

- On the other hand, high profile is suitable for applications which demand high compression efficiency and has a relaxed system latency requirements such as broadcasting. It supports progressive and interlaced coding as well as “adaptive coding” of progressive/interlaced content at both picture level (PAFF) and MB level (MBAFF). It also supports encoding of B-frames in addition to I, P-frames and both CAVLC and CABAC encoding options are provided. The error resilient tools such as arbitrary slice ordering (ASO), redundant slices and flexible macro block ordering are not supported. In summary, this profile is suitable for off-line content encoding and is used in mass storage and local media playback.

The level values specified in the standard ranges from 1.0 to 5.2 and provides a good amount of flexibility for various tools including bit rate and spatial resolutions. In terms of spatial resolution, level 1.0 supports Quarter Common Intermediate Format (QCIF) [44] and Level 5.2 supports Ultra High Definition (UHD) format. Similarly in terms of bit rate, level 1.0 supports 64 Kilobits per second and level 5.2 supports few hundred megabits per second. Thus the standard could be used in a wide variety of applications.

### **3.1.2. Encoder operation**

The high-level encoding architecture of the H.264/AVC encoder is shown in Fig.3.1.



© 2007, IEEE

**Figure 3.1:** High-level coding architecture of the H.264/AVC encoder [37]

Following ordered steps are executed on the H.264/AVC encoder,

- Every frame in the video sequence is split in to a number of slices and in multi-processor environments, slices could be encoded in parallel.
- Every slice is classified as Intra (I), Predictive (P) or Bi-predictive (B) based on the video content and available bitrate.
- The slices will be further divided in to a number of Macro blocks (MB) with the MB size as 16x16

**Intra slice encoding process:**

- In case of intra slice, all the MBs will be encoded as intra and there are multiple MB type options available such as Intra\_16x16, Intra\_8x8 and Intra\_4x4.
- Intra\_16x16 is suitable for uniform content and Intra\_8x8, Intra\_4x4 are useful to capture small variations in the frame.
- There are four modes for Intra\_16x16 based on the availability of neighboring pixels and similarly Intra\_4x4, Intra8x8 supports nine modes based on the availability of neighboring pixels.
- Encoder evaluates all the MB types along with the mode options and choose the best among them based on the rate-distortion performance.
- Once the best MB type and mode are chosen, the predicted pixels will be subtracted from the original pixels and the resultant residual pixels will be transformed by using Integer transform (IT) which is

based on Discrete Cosine Transform (DCT) [45]. The kernel size of the integer transform is either 4x4 (or) 8x8 and it could be noted that 8x8 transform is available only in high profile.

- The transformed pixels will be subject to quantization process and the purpose of the quantization is to remove the high frequency components, which has a small but non-zero magnitude.
- The transformed, quantized pixels will be encoded by using either CAVLC (or) CABAC encoding, with header fields encoded by using simple VLC (variable length coding) coding.
- Also, the transformed, quantized pixels will be inverse quantized, inverse transformed and added with the predicted pixels to produce the motion compensated signal. The reconstructed, motion compensated pixels will be used as prediction signal for the neighboring MBs.

#### **Inter slice encoding process:**

- The slice could be encoded either as predictive slice (P-Slice) or Bi-predictive slice with some of the MBs in the slice could be encoded as Intra.
- The MB types available for the predictive MBs (P-MBs) are P\_SKIP, P\_16x16, P\_16x8, P\_8x16, P\_8x8, P\_8x4, P\_4x8, and P\_4x4. P-MBs will have a single reference picture list and each sub-MB (8x8) can have its own reference index and each 4x4 will have a unique motion vector.
- P\_SKIP is a special case, where the motion and reference index information will be copied from the previously coded co-located MBs and this MB type will be useful for static content across the frames.
- As mentioned earlier, intra MBs could also be coded as part of inter slices, hence the encoder has to evaluate both the P-MB type related options as well as intra MB type options for a given P-MB.
- Once the MB type is chosen, the rest of the encoding process is same as that of Intra MBs.
- In case of B-MBs, there will be two reference picture lists and the options will be on the same lines as that of P-MBs. The special modes include B\_SKIP and B\_DIRECT, where B\_SKIP will be similar to P\_SKIP. But in the case of B\_DIRECT, the motion vectors will be inferred from either the spatial (Spatial B\_DIRECT) / temporal (Temporal B\_DIRECT) neighbors.
- Again, the encoder has to evaluate the I\_MB options in addition to the B\_MB options for a given B\_MB and the rest of the encoding process is same as that of I\_MB.

#### **3.1.3. Multiple reference pictures and Decoded Picture Buffer**

In earlier video coding standards such as H.263 [27], MPEG-4 Visual [26], there will be only one reference picture, which is the previously encoded frame and H.264/AVC introduced the concept of multiple reference pictures for the case of motion compensated prediction. Multiple reference pictures were originally proposed by Wiegand et al. in the case of H.263 and experimental results shows that approximately 20% reduction in bit rate

could be achieved by using the method [46]. Some earlier experimentation by Puri et al. showed that around 10% reduction in bit rate could be achieved in the case of H.264/AVC [47].

Another important aspect introduced in H.264/AVC is the de-coupling of decoding order from the display order. In earlier standards, except for the B-pictures, the decoded picture will be immediately sent to display. But in H.264/AVC, the display of the decoded picture could be delayed, irrespective of the picture type. This will be extremely useful to achieve temporal scalability and to generate sub sequences. In standards such as MPEG-2, sub sequences will be generated by simply dropping the B-pictures [48], as B-pictures will not be used for reference, there will not be any impact on the decoding process. But in H.264/AVC, this method will no longer work as even B-pictures could be used for reference and keeping the information as to whether the picture will be used for future reference or not will require deep inspection of the packets at the network element. In order overcome this, an additional field termed as 'nal\_ref\_idc' is introduced in the NAL header and this field will convey the information as to whether this picture will be used as reference for future pictures or not. Hence, sub sequences could be generated by simply inspecting this flag and same is true for the case of temporal scalability as well.

Since there are multiple reference pictures, in order to store them a decoded picture buffer (DPB) is introduced. The decoded pictures which has a valid non\_ref\_idc, will be sent to DPB and stored for reference. Before encoding a picture, the reference picture lists will be formed from the DPB, where List0 contains the decoded, displayed pictures and List1 contains decoded, not yet displayed pictures. It could be noted that Predictive pictures will have only List0 for reference and Bi-predictive pictures can refer to both List0 and List1. The encoder will evaluate the reference pictures and pick the best among the available for a given sub-MB (8x8). The chosen reference picture index will be encoded in the bitstream along with the reference picture list information.

There are two types of reference pictures namely short term and long term. Short term reference pictures are useful when there is a linear motion across frames and long term reference pictures are useful in scenarios where there is repeated frame contents in the video sequence after periodic interval. Short term reference pictures are identified by the frame\_num and long term reference pictures are identified by the long term picture number. During the construction of reference picture list, first the short term reference pictures will be placed in the ascending order of frame\_num followed by the long term reference pictures in the descending order of long term picture number.

#### **3.1.4. Memory management process**

As mentioned in the last section, the decoded frames will be stored in the DPB for future reference. Also, short term reference pictures are identified by the "frame\_num" field and long term reference pictures are identified by

the "Long\_term\_pic\_num" field. After decoding a frame, the frame will be marked as "used for reference" and stored in to the DPB. Before starting the encoding/decoding process, reference picture list is constructed from DPB using the following protocols,

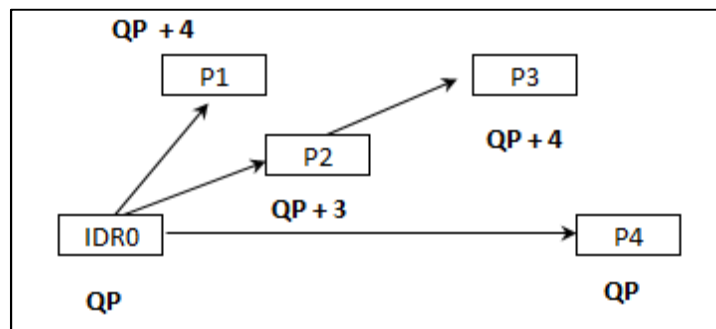
- Short term reference pictures will be placed first in the list followed by long term reference pictures
- Short term reference pictures should have "frame\_num" which is lesser than the "frame\_num" of the current frame
- Long term reference pictures should have "Long\_term\_pic\_num" which is greater than the "Long\_term\_pic\_num" of the current frame
- There should be at least one short term reference picture in the reference picture list
- Short term reference pictures will be arranged in the ascending order based on the "frame\_num" value and Long term reference pictures will be arranged in the descending order based on the "Long\_term\_pic\_num" value

The above process is termed as "default reference picture list construction process" and the same is executed on both encoder and decoder before encoding/decoding of the frame. After the "default" reference picture list is constructed, as an optional step, the encoder can re-arrange the reference picture list by using the "reference picture list re-ordering" (RPLR) commands. In general, encoders will execute algorithms which will find the "best matching" reference frames from the contents of the reference picture list and in case if the best matching reference pictures are not present in the initial part of the "default" list, list will be re-arranged by using the RPLR commands. The advantage of the process is that the "compute intensive" operations such as motion estimation could be limited to sub set of the reference picture list (initial few reference pictures) and the consumption of bits w.r.t the encoding of reference picture index could also be minimized.

Once the picture is encoded/decoded, there are two additional "memory management" options available namely "memory management control operations" (MMCO) and Sliding window process. In the process of MMCO, explicit memory management is done to the contents of the reference picture list. For example, a short term reference picture could be converted in to a long term reference picture, vice-versa and a picture could be marked as "unused for reference" etc. On the other hand, a sliding window process works on a "First in First Out" (FIFO) model and when the decoded picture buffer is full, the short term reference picture with the lowest "frame\_num" value will be marked as "unused for reference" and pushed out of DPB. It could be worth mentioning that the "instantaneous decoding refresh" (IDR) picture will implicitly mark all the short term and long term reference pictures as "unused for reference" and typically this will be used as the first picture of the encoding process for the given video sequence.

### 3.1.5. Prediction structures

The prediction structure plays an important role in many aspects such as compression efficiency, structural latency, memory requirements etc. and it is vital to balance these aspects. As mentioned in the previous paragraphs, unlike other video coding standards, H.264/AVC introduced the concept of using any coded picture as reference including the B pictures. It has been shown that the H.264/AVC sub sequences obtained by marking some of the P/B pictures as unused for reference (also known as disposable p/b pictures), provides improved compression efficiency and also enables the extraction of scalable bitstreams [48]. It was shown that compression efficiency obtained by disposable frames based hierarchical B coding is better than that of motion compensated temporal filtering (MCTF) with reduced encoder complexity, as the MCTF based encoder needs to perform additional update steps [49]. Disposable frames based hierarchical coding also offers a number of other advantages such as reduced encoder complexity as the reconstruction path (inverse quantization, inverse transformation, motion compensation and in-loop filter) could be skipped for the highest temporal layer pictures [50]. It has also been shown that the coding efficiency could be further improved by encoding temporal layers at different quantization parameter (QP cascading scheme), with lowest temporal layer encoded at a QP which is less than that of the next temporal layer which in turn was encoded at a lower QP compared to its next temporal layer and so on. An example prediction structure is shown in Fig.3.2.



**Figure 3.2:** Prediction structure with quantization level set corresponding to the layer. Base layer will be encoded using lowest QP.

### 3.2 A brief overview of the H.264 MVC standard

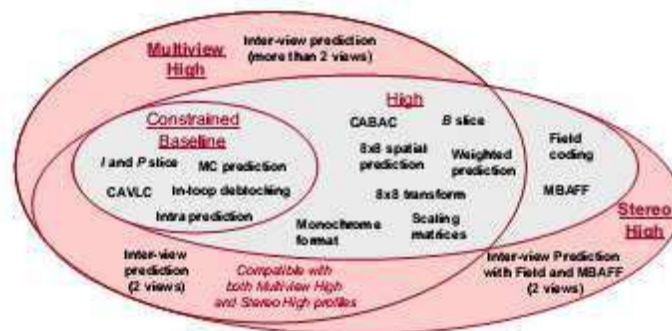
H.264 multi view video coding (H.264 MVC) is based on H.264/AVC specification, but is used for encoding of multi view video data. As mentioned in Chapter.1, multiple synchronized cameras capture the same scene and fed as an input to the MVC encoder. The easiest way to compress the video data from all the cameras is to encode individual camera data by using a separate H.264/AVC encoder and multiplex the individual resultant



H.264/AVC bitstreams to form the final H.264 MVC bitstream. This method is called simulcast and is using only the temporal neighbors for prediction, view neighbors (data captured by other cameras) are not used in the prediction process. On the other hand, inter-view prediction is expected to improve the compression efficiency as the same scene was captured by a number of cameras and hence there will be similarities between the video content. But having more number of view components as predictors will increase the encoder complexity as it has to evaluate a number of interview predictors in addition to the temporal predictors to find the best suitable option. Also, this option will increase the memory requirements as the view components corresponding to various cameras, captured at different time instances needs to be stored. In order to study the usage of inter-view prediction, MPEG issued call for proposals and studies showed that MVC can achieve significant results when compared to simulcast [51].

Studies also showed that in most of the cases, temporal predictors provide better representation of the scene content when compared to interview predictors [52]. But there are certain scenarios in which the inter-view prediction provides a better representation and the scenarios include camera arrangement in which the distance between the cameras is small and the capture frame rate is very less. Hence in summary, it is worth evaluating the data from other cameras in addition to the temporal content from the same camera. Thus, MVC specification contains both temporal prediction and inter-view prediction, but to reduce the number of inter-view predictors, data from the other cameras captured at the same time instant only is considered.

### 3.2.1. Profiles and level



© 2011, IEEE

**Figure 3.3:** An illustration of MVC profiles, consisting of the multi view High and Stereo High profiles [11]

H.264 MVC supports two profiles namely Stereo High Profile and Multiview High Profile. Fig.3.3 illustrates the supported coding tools with respect to various profiles. Stereo High Profile (SHP) is capable of encoding two views and the base view could be encoded using constrained baseline profile (or) high profile. In case if the base

view is encoded using constrained baseline profile, interlaced coding tools (MBAFF, PAFF) are not supported. Instead, if the base view is encoded using high profile, then interlaced coding tools are supported. This profile is intended for stereo video, as in many application areas stereo video encoding is considered as the starting point towards multi view video processing.

Multiview High Profile Supports a maximum of 1024 views and is suitable for broadcast applications with minimal spacing between the cameras. Similar to Stereo High Profile, the base view could be encoded using constrained baseline profile (or) high profile. When the multiview high profile (MHP) is used for the encoding of stereoscopic video, there is an ambiguity between the MHP and SHP profiles and in order to resolve the ambiguity, a flag will be set to indicate the compatibility. Also in both the profiles, there is a provision exists to turn-off inter-view prediction.

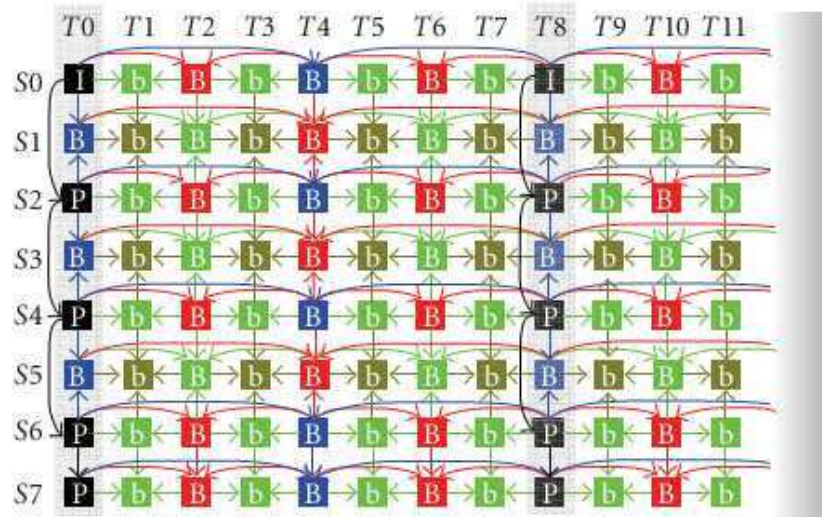
As mentioned in Sec.3.1, levels add details and imposes constraints on the supported features of the bitstream. Level values are enhanced for the case of H.264 MVC, for example in H.264/AVC, level 5.1 supports encoding of 1920x1080 (1080p) video at 30 frames per second. In MVC case, this translated in to encoding of stereo 1080p at 30 frames per second. Also in the case of MVC, multiple level values could be signaled in the header, with each level value correspond to a sub-sequence.

### 3.2.2. MVC Prediction Structures

There are a number of prediction structures proposed in the literature and Fig.3.4 shows a structure developed by Fraunhofer HHI for the case of a 1D camera arrangement (linear or arc) [34]. ‘SX’ in the figure represents the camera numbers, ‘TX’ denotes the instances in time. Similarly, ‘B’ represents the ‘bi-predictive’ view components which are used as reference for the other view components and ‘b’ stands for ‘bi-predictive’ view components which are not used for reference. ‘b’ view components are also termed as ‘disposable’ view components, as these view components will be discarded after the display and will not be stored in the DPB. The proposed MVC prediction structure has several advantages,

- The disposable view components helps in improving the encoder/decoder speed as the reconstruction path (inverse quantization, inverse transformation, motion compensated prediction, in-loop filters) could be turned off for these view components.
- Compression efficiency could be further improved by “QP cascading scheme” as outlined in Sec.3.1.5. The ‘b’ view components will be encoded at a higher QP when compared to ‘B’ view components as these view components will not be used for reference and there will not be any propagation of quantization error.

- Time-first coding order is maintained, where the view components captured at the same time instant ‘T0’ will be encoded first followed by the coding of view components captured at time instant ‘T1’ (T1 > T0).
- Intra view components are inserted at periodic time intervals, this will be helpful in scenarios such as random access and error resilience. For example, if there is a packet loss when transmitting the compressed bitstream from encoder to decoder, the decoder will be in synchronization with the encoder from the synchronization point (where base view is encoded as Intra).
- Though the “synchronization point” interval is less, only the base view in the synchronization point is coded as intra and other view components are coded as Inter (either predictive or bi-predictive). Hence it achieves a good trade-off between compression efficiency and error-resilience. Moreover, the prediction structure utilizes both temporal view components and inter-view components for prediction, as it has been observed that camera arrangement plays a major role in selecting the best among temporal and inter-view predictors [54].
- Additional cameras could be added and existing camera arrangement could be altered seamlessly, which will be useful for the media aware network element to extract the required packets. Though the MVC prediction structure shown in Fig.3.4 could serve a number of devices with various “view requirements”, it could not produce sub sequences within the same camera data and the possible related improvements are discussed in Chapter.5.



© 2007, IEEE

**Figure 3.4:** State-of-the art prediction structure for H.264 MVC [53]

### 3.2.3. Encoder operation

There are no MB level coding tools (apart from the existing H.264/AVC coding tools) introduced in H.264 MVC. To enable inter-view prediction in addition to the existing temporal prediction, the encoder operates in such a way that the base view will be encoded first followed by the enhancement views. For the first picture of the sequence, the base view component will be encoded as IDR and the enhancement view components will be coded as predictive (or) bi-predictive based on the available inter-view neighbors. H.264 MVC design mandates individual DPBs for every view, hence for the MVC decoder with two views there will be two separate decoded picture buffers. The ‘default’ reference picture list construction process for the base view component follows the procedure outlined for H.264/AVC and for the case of H.264 MVC it works as follows,

- Short-term ‘temporal’ reference view components with ‘frame\_num’ value lesser than the ‘frame\_num’ of the current view component *and* has the ‘view\_id’ same as that of current view component are placed first in the reference picture list.
- Long-term ‘temporal’ reference view components with ‘long\_term\_frame\_num’ value greater than the ‘long\_term\_frame\_num’ of the current view component *and* has the ‘view\_id’ same as that of current view component are placed next in the reference picture list.
- View components from other cameras with ‘view\_id’ value less than the ‘view\_id’ of the current view component *and* captured at the same time instant ‘t’ are placed next in the reference picture list.
- At this stage, the reference picture list may have view components at indices greater than the ‘num\_reference\_pictures’ coded in the parameter sets. In these circumstances, the view components located in the indices greater than the ‘num\_reference\_pictures’ will be discarded ‘irrespective’ of the ‘view\_id’ of the reference view component.

Once the ‘default’ reference picture list is constructed, the reference picture list re-ordering (RPLR) process is invoked and according to H.264 MVC specification the RPLR commands need to be signaled for individual view components and same is applicable even for the memory management control operations as well. There are some disadvantages of the existing H.264 specification and in Chapter.5, some of the issues are illustrated, mechanisms to address the issues have been detailed along with extensive simulation results.

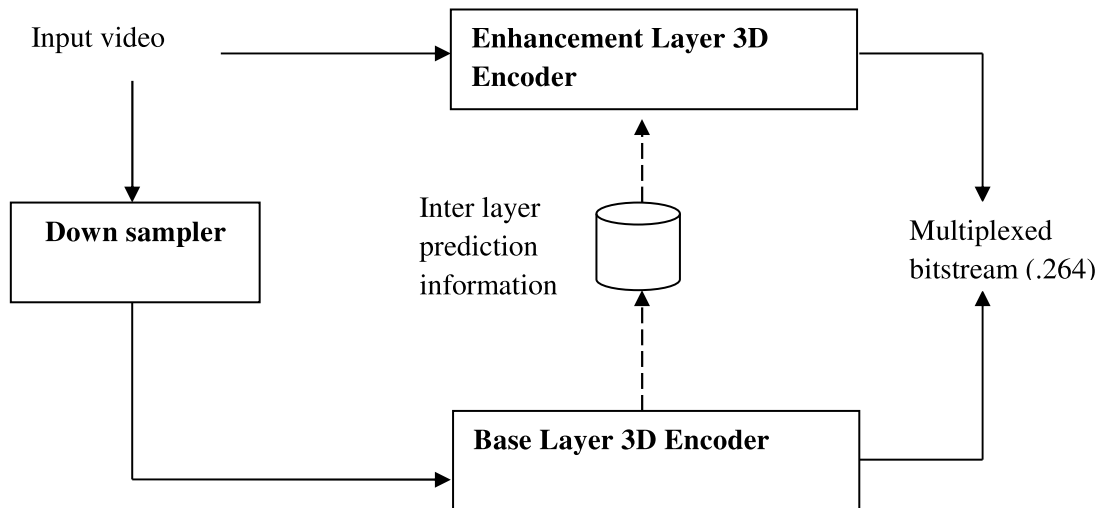
## Chapter 4

### Scalable Multi View Coding Scheme: Design aspects

This chapter provides a brief introduction to the design aspects of the proposed H.264 Multi view video coding specification based scalable multi view coding scheme. The intent of the chapter is to provide an incomprehensive overview of the scalable multi view coding scheme including the structure of the bitstream, extraction and adaptation of bitstreams, software architecture, test procedure and the test sequences used for the evaluation and benchmarking process.

#### 4.1. Layered architecture

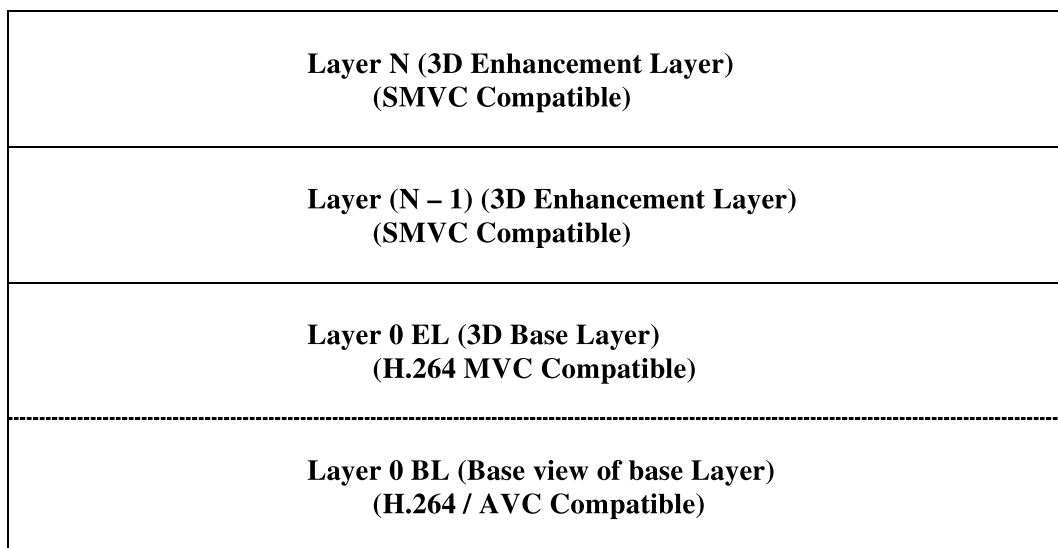
The layered architecture of the H.264 MVC based scalable multi-view coding scheme is shown in Fig.4.1. The architecture is based on time-first coding order to utilize the file format design of MVC. In the time-first coding order, a particular frame 'T' corresponds to all the views of the base layer will be encoded first, followed by the encoding of the same frame 'T' corresponds to all the views in subsequent layers. On the other hand, in a view-first coding order, all the frames of a particular view 'V' will be encoded first, followed by the encoding of the frames at other views. Thus, time-first coding order is essential to achieve optimal buffer management at the decoder and to reduce the end-to-end latency of the system as well.



**Figure 4.1:** Layered architecture of the scalable multi view coding encoder with two layers.

The base view encoder in the base layer evaluates the temporal prediction tools (MB level prediction tools) as specified in the H.264/AVC specification. The enhancement view encoders of the base layer evaluate both

temporal and inter-view prediction tools as specified in the H.264 MVC specification. The temporal and inter-view predictions are collectively referred to as intra-layer prediction tools. The enhancement layer encoder (both base and enhancement views) utilizes information from the reference layers, which is termed as inter-layer prediction tools in addition to the intra-layer prediction tools(temporal and inter-view predictions). The encoders operates in tandem, i.e. the base layer encodes multiple frames with frame number 'N' corresponding to multiple views in the base layer, followed by the enhancement layer encoder which also encodes multiple frames with frame number 'N' corresponding to multiple views with respect to the enhancement layers, but at a different bit rate (coarse grain scalability) / different resolution (spatial scalability). The NAL unit (NALU) packets from all the encoders are multiplexed in to a single output stream. More than one view support (view scalability) and multiple frame rate (temporal scalability) support will be provided at every layer. It could be noted that though only two layers are depicted in the Fig.4.1, it is possible to have more than one enhancement layer and every layer contains a number of views (base view and a number of enhancement views). The base layer encoder stores information such as residual pixels, reconstructed pixels, motion information and partition information, which will be used as inter-layer prediction data at the enhancement layers. For the case of spatial scalability, the input video will be corresponding to the highest enhancement layer and will be down sampled before providing the same as the input to subsequent layers.



**Figure 4.2:** Packet arrangement of the output stream of the scalable multi view coding encoder. Base view of the base layer is AVC compatible 2D view and rest of all layers is MVC compliant.

The base layer encoder produces streams which are compatible with the H.264/MVC specification (for legacy 3D receivers), which also implies that the base view of the base layer is backward compatible with H.264/AVC specification (for legacy 2D receivers). The enhancement layer encoder(s) produces bitstreams which are

compatible with the scalable multi view coding specification. This layered arrangement of the bitstream as shown in Fig.4.2 ensures that the media aware network element could extract the required NAL packets with ease.

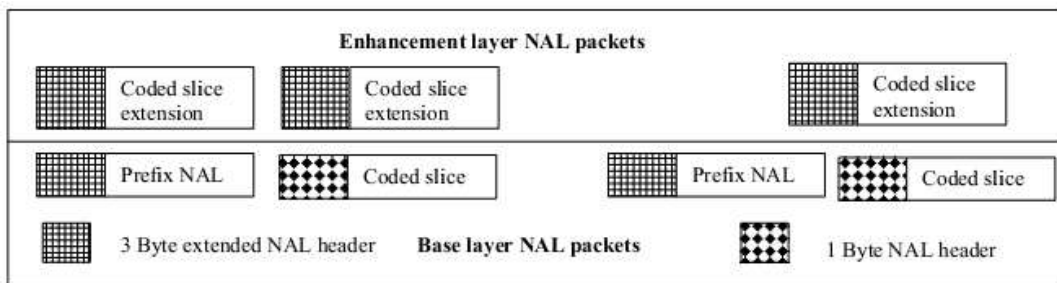
## 4.2. Extraction and adaptation of bitstreams

As mentioned in Sec.3.2, MVC already supports temporal scalability and view scalability and these values are signaled as 'view\_id' and 'temporal\_id' in the NAL header. Signaling these flags in the NAL header has several advantages, MANE can serve several end points with various 'view requirements' using a single encoded bitstream, when there is a reduction in available bandwidth MANE can drop selective pictures (with view components which has large temporal\_id). Though H.264 MVC specification does not place any restrictions on the 'temporal\_id' value for a given view component (as an example, view components captured by different cameras but same time instant can have different 'temporal\_id'), it is an implicit requirement that the view components captured at the 'same time instant' will have same 'temporal\_id'. Also, a view component with 'view\_id'  $V_x$  will be used as a reference for view components with view\_id  $V_y$ , where  $V_y$  is greater than  $V_x$ . this process will ensure that when part of the encoded bitstream is extracted at MANE, all the other dependent view components are encapsulated in the extracted sub bitstream. Same requirement is applicable for temporal\_id as well and it could also be noted that though temporal scalability is supported by H.264 MVC, there are possible improvements to achieve more rate-distortion performance and the same are discussed in Chapter.5.

In the scalable multi view video coding scheme, in addition to the temporal and view scalability, spatial and bit-rate scalabilities will also be supported and for this purpose an additional flag called layer\_id is introduced in the NAL header. Since there is a reserved bit already exists in the NAL header as shown in Fig.4.3, this bit will be re-named as layer\_id and used in the SMVC specification. This flag will be zero for the base layer NAL units (MVC compatible layer) and will be set to one for the enhancement layer NAL units (SMVC compatible layers). Thus, the enhancement layer NAL units will be treated on par with the enhancement view NAL units of MVC, where the NAL header is an extended NAL header of 3 bytes. Thus, in summary the packet arrangement in the proposed SMVC compression scheme works as follows; base view packets of base layer, which is H.264/AVC compatible will be prefixed using 1-byte NAL header, enhancement views of the base layer (MVC compatible) and enhancement layer packets (SMVC compatible) will be prefixed by a 3-Byte NAL header as shown in Fig.4.4.

nal_unit_header_mvc_extension(){	C	Description
<b>non_idr_flag</b>	All	u(1)
<b>priority_id</b>	All	u(6)
<b>view_id</b>	All	u(10)
<b>temporal_id</b>	All	u(3)
<b>anchor_pic_flag</b>	All	u(1)
<b>inter_view_flag</b>	All	u(1)
<b>reserved_one_bit</b>	All	u(1)
}		

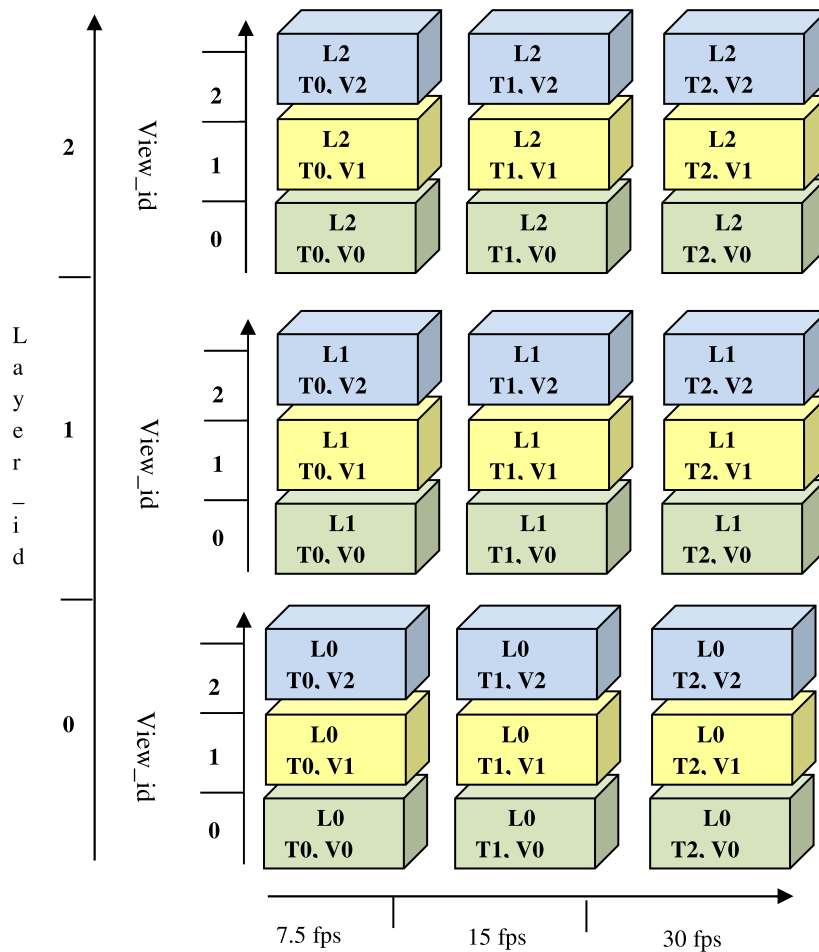
**Figure 4.3:** NAL header of H.264/MVC complaint stream. Solid box indicates the reserved bit.



**Figure 4.4:** Packet arrangement of the SMVC scheme

Hence in summary, the output bitstream will be made up of a number of layers as shown in Fig.4.5 and the multiplexed bitstream will be sent to gateway / server. For example, in the bitstream arrangement shown in Fig.4.5, there are three layers (L0, L1 and L2), each contains three views (v0, v1 and v2) and three temporal levels (T0, T1 and T2). The H.264 MVC compatible base layer (L0) can support a resolution of 720p (1280 x 720) at 4 Mbps and the enhancement layer (L1) can support a resolution of 1080p (1920 x 1080) at 8 Mbps and finally the highest enhancement layer (L2) can support a resolution of 1080p (1920 x 1080) at 12 Mbps. Every layer contains three views and three temporal levels (7.5 fps, 15 fps and 30 fps) and based on the decoder needs, some of the views / frames could be discarded to meet the receiver requirements. The media gateway / server / MANE receive the bitstream and based on the capability of the receiver, the server will extract the appropriate layer's packets and forward the same to the receiver. For example, if there is a legacy 3D receiver connected to the network, which is capable of 720p, then the base layer(L0) packets will be sent and similarly for a legacy 2D receiver, which is capable of 720p, the packets correspond to the base view of base layer will be extracted and sent. For a SMVC receiver, which is capable of 1080p @ 8 Mbps, 15 fps, both L0 and L1 packets, with temporal\_id as T0 and T1 will be sent and so on and so forth.

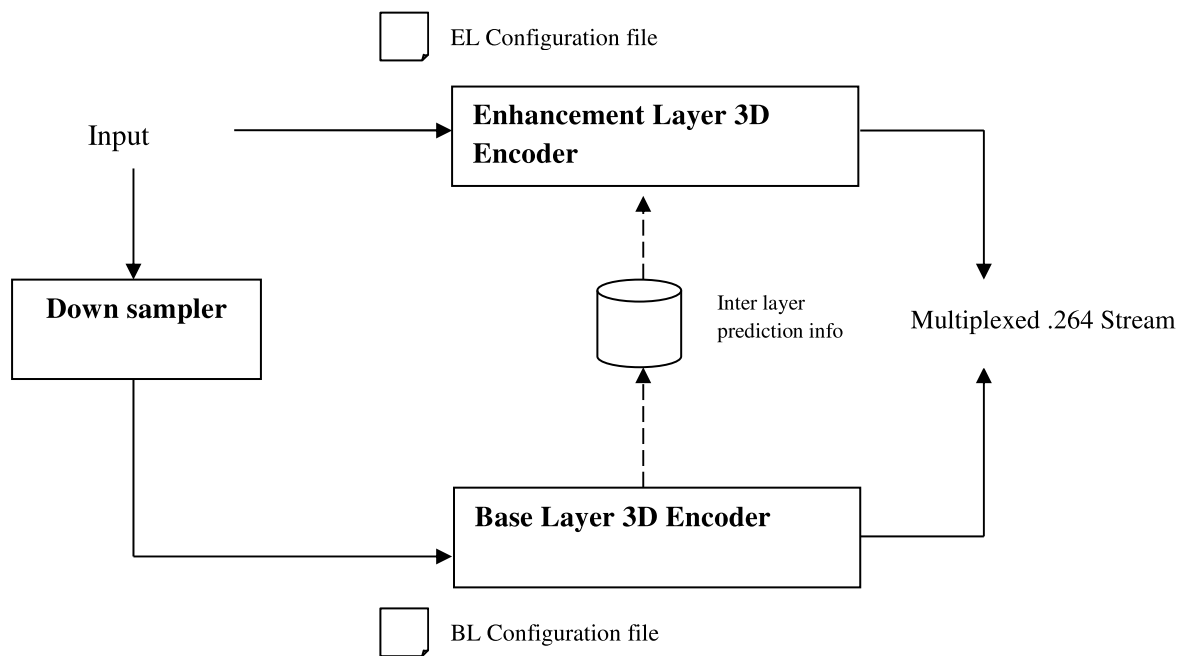




**Figure 4.5:** Packet arrangement of the SMVC scheme with two spatial layers, one quality layer each with three temporal and three view layers

### 4.3. Software overview

The JM reference software provided by Joint Video Team (JVT) [43] provides support for the encoding/decoding of stereoscopic video and the same has been used for the software development of scalable multi view coding scheme. Though multi view video support is available with the JMVC reference software [55], the support for the software has been discontinued by JVT and hence JM software is chosen as the starting point. First, the temporal scalability support is added to the JM reference software and the number of supported layers has been increased to more than one. The software architecture of JM reference software used for the SMVC software development is shown in Fig.4.6.



**Figure 4.6:** Overview of SMVC software using JM reference software

The software is controlled by a set of configuration files, where the number of configuration files is equal to the number of layers to be encoded. For example, if there are three layers to be encoded, then the number of enhancement layers will be specified in the base layer configuration file and each enhancement layer data will be present in separate configuration files. This arrangement ensures that the encoding tools could be turned on/off at every layer level and the path of the enhancement layer configuration files are specified in the base layer configuration file such that the encoder can validate the number of configured layers with the configuration files at the base layer level itself. In order to simulate the gateway/server, which has the ability to perform operations such as bitstream adaptation and extraction, the bitstream extractor software [43] provided by the JMVC reference software is used. The software works on the basis of operation point based extraction and has been enhanced to have the support for temporal\_id/view\_id/layer\_id based packet extraction. Thus, the software can be configured through command line to accept the inputs such as target\_temporal\_id, target\_view\_id, target\_layer\_id and only the relevant packets will be extracted.

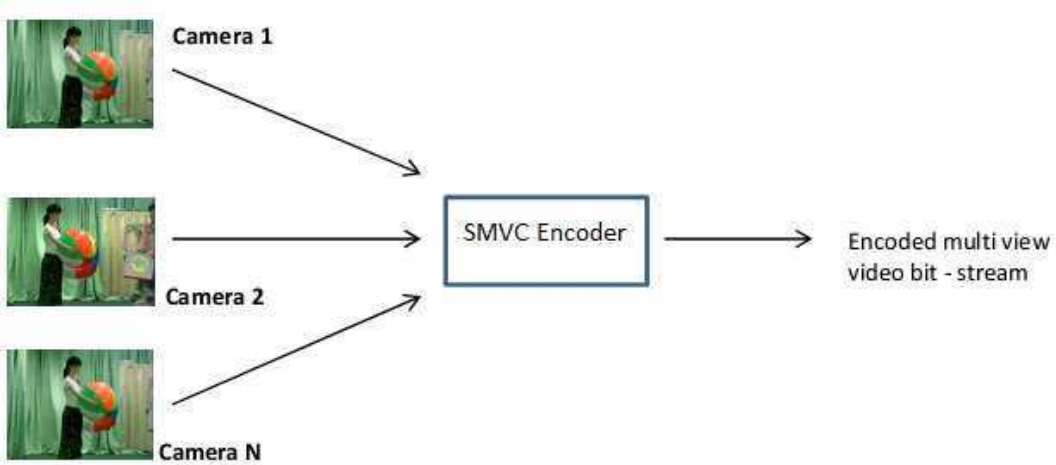
#### 4.4. Test procedure and Test sequences

The test procedure is depicted in Fig.4.7 to Fig.4.9, where the input video is encoded using SMVC encoder (JM reference software) and the resultant bitstream will be provided as an input to the bitstream extractor software (JMVC bitstream extractor software) to extract the required packets. The extracted packets (NAL units) will be sent to the SMVC decoder (JM reference software decoder) for final play back. The bitstream extractor will be configured with different inputs to simulate the different receivers and the extracted packets will be sent to the SMVC decoder (JM reference software decoder). There are three options provided in the configuration file of the enhancement layers,

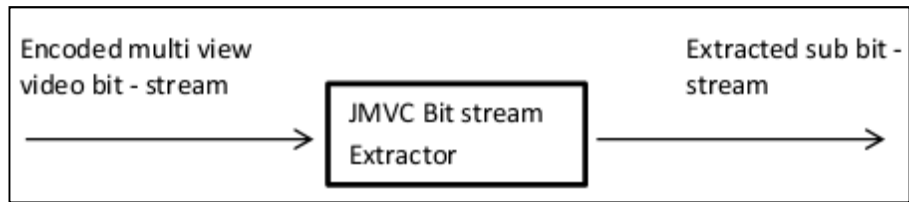
- **Simulcast option:** The encoder will evaluate only the intra-layer prediction tool (temporal and inter-view prediction tools) and the inter-layer prediction tools will not be evaluated. In other words, the encoders will be H.264 MVC compliant and encode the videos independently.
- **Always inter layer prediction option:** The encoder will evaluate the inter-layer prediction tools only and select the best among them, without evaluating the intra-layer prediction tools. It could be noted that this mode is used for evaluation purposes only.
- **Adaptive inter layer prediction option:** The encoder will evaluate both inter-layer and intra-layer prediction tools and select the best among them.

The mode selection in all the above methods involves rate-distortion optimization [56] which was introduced in low-bit rate coding specifications such as H.263 [57]. In the rate-distortion scheme, all the modes will be evaluated based on both the bits required for encoding the mode (including the header information) and the resultant distortion.

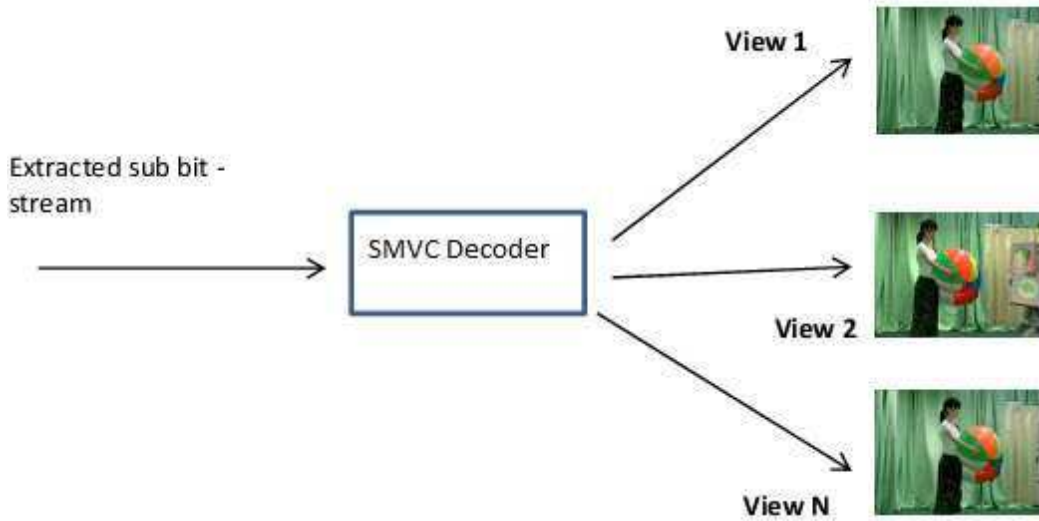
The test vectors include various resolution input sequences such as VGA captured using [58] up to eight cameras and XGA resolution camera sequences [59] using up to 100 moving cameras. To ease the test procedure, only two views are used in the evaluation process and a limited set of quantization parameters (QP) have been used to measure the efficiency of the method at different bit rates.



**Figure 4.7:** Encoding of multiple view videos using SMVC encoder



**Figure 4.8:** Extraction of required packets using JMVC bitstream extractor



**Figure 4.9:** Decoding of SMVC bitstream using SMVC decoder (JM reference software)

## **Chapter 5**

### **Temporal scalability**

This chapter explains the modifications proposed to the existing temporal scalability in H.264 Multi view coding. Though temporal scalability is supported by H.264 MVC, there are couple of possible improvements as listed below.

- The temporal scalability requirements could not be met to the fullest by using the current state-of-the-art prediction structure for MVC, which is based on "disposable view components". A new state-of-the-art prediction structure, which can generate a fully scalable bitstream, has been proposed in Sec.5.1 of the chapter.
- In the current H.264 MVC specification, the temporal\_id information is not used in the reference picture list construction process. Sec.5.2 of the chapter proposes to use the temporal\_id information in the list construction process for a better rate-distortion optimization performance compared to existing H.264 MVC.

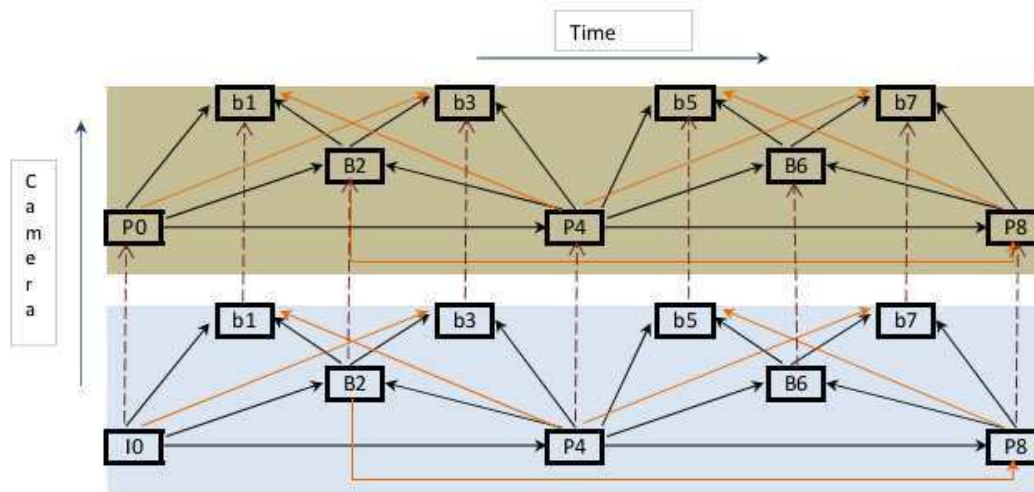
It could be noted that second section of the chapter proposes some modifications to the H.264 MVC specification text to achieve better rate-distortion performance. This chapter summarizes the publications [J2] and [P1].

#### **5.1. Prediction structures for H.264 MVC**

##### **5.1.1. Disposable view components based hierarchical coding**

Unlike other video coding standards, H.264/AVC introduced the concept of using B frames as reference and the idea has been extended to H.264 MVC as well. It has been shown that sub sequences obtained by marking some of the P/B pictures as unused for reference (also known as disposable p/b frames), provides improved compression efficiency and also enables the extraction of scalable bitstreams [48]. It was shown that compression efficiency obtained by disposable frames based hierarchical B coding is better than that of motion compensated temporal filtering (MCTF) with reduced encoder complexity, as the MCTF based encoder needs to perform additional update steps [49]. Disposable frames based hierarchical coding also offers a number of other advantages such as reduced encoder complexity as the reconstruction path (inverse quantization, inverse transformation motion compensation and in-loop filter) could be skipped for the highest temporal layer pictures [50], thereby increasing the encoder speed. It has also been shown that the coding efficiency could be further improved by encoding temporal layers at different quantization parameter (QP cascading scheme), with lowest temporal layer encoded at a QP which is less than that of the next temporal layer which in turn was encoded at a

lower QP compared to its next temporal layer and so on. The concept of “disposable frames based” hierarchical coding has been derived from H.264/AVC to H.264 MVC as well [53] and one such prediction structure corresponding to stereoscopic video is shown in Fig.5.1. It could be noted that this prediction structure is used in H.264/AVC reference software JM [60] which also supports encoding of stereoscopic video using H.264 MVC.



**Figure 5.1:** Disposable B frames based hierarchical coding. Dashed lines indicate inter-view prediction and solid lines represents temporal prediction. The top-box represents view components from right view camera and the bottom box depicts the view components from left view camera.

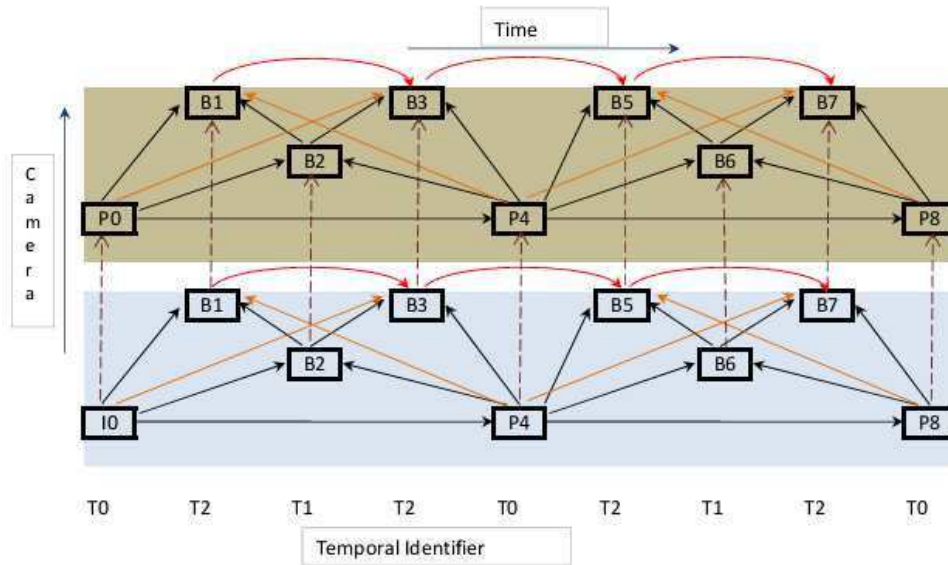
In the depicted prediction structure, “I0” represents the start of the video sequence and “P0” represents the anchor view component in right view. Number of reference view components used is three. “b” view components represents the disposable view components, by which bit rate adaptation could be achieved in the media aware network element (MANE) such as media gateway and routers. When the bandwidth of the existing network (such as IPTV connected via broadband network) drops down, MANE could drop some view components to cope with the problem. It could be noted that the frame rate will be reduced by half after removing view components belongs to a layer. For example, if the video is encoded at 60 fps, then removing all “b” view components from the stream would produce a video at 30 fps which requires less bit rate to transmit, when compared to the full video encoded at 60 fps. But the disposable view components based hierarchical coding has limitations, as it could be seen that, when the subsequent temporal layer view components (which consists of view components such as “B2, B6”) are removed, the video quality will have a serious impact, as “B2” is used as a reference for “P8”, which belongs to the lowest temporal layer. It is worth to mention that the prediction structure could be modified in such a way that this scenario (of layer (N+1)<sup>th</sup> view component is used as reference for a view component which is present in layer N) will not occur, but this will reduce the compression efficiency as well, as the number of temporal predictors will be reduced. Nevertheless, there is a problem in the identification of the layer (at which a view component belongs to) itself at MANE, as it relies only upon the disposable property

(signaled using `nal_ref_idc` in the NAL header) of the packet and in the example shown above, apart from the highest temporal layer, all the other layers' view components are coded with non-zero `nal_ref_idc`. This makes it impossible for the MANE to identify the view components which belongs to layer '(N- 1)', such that the frame rate would be reduced by half by removing packets present in layer (N-1).

### **5.1.2. Temporal identifier based hierarchical coding**

Temporal identifier based temporal scalability was introduced in H.264 scalable video coding (SVC) [31] with the NAL header extended from one byte in H.264/AVC to three bytes in SVC, which also includes the temporal identifier (`temporal_id`) information. H.264 MVC derived a number of system level information from H.264 SVC, one of them being `temporal_id` information as part of the extended NAL [61]. The prediction structure for `temporal_id` based hierarchical coding is shown in Fig.5.2, which contain three temporal layers ("T0", "T1" and "T2"). The restriction imposed by the specification is that, `temporal_id` of the reference view components should be less than or equal to 'N' for view components with `temporal_id` "TN". This enables the MANE to extract the view components of layer (N+1) to reduce the bit rate as the layer (N+1) will be used as reference only for view components with `temporal_id` greater than or equal to (N+1).

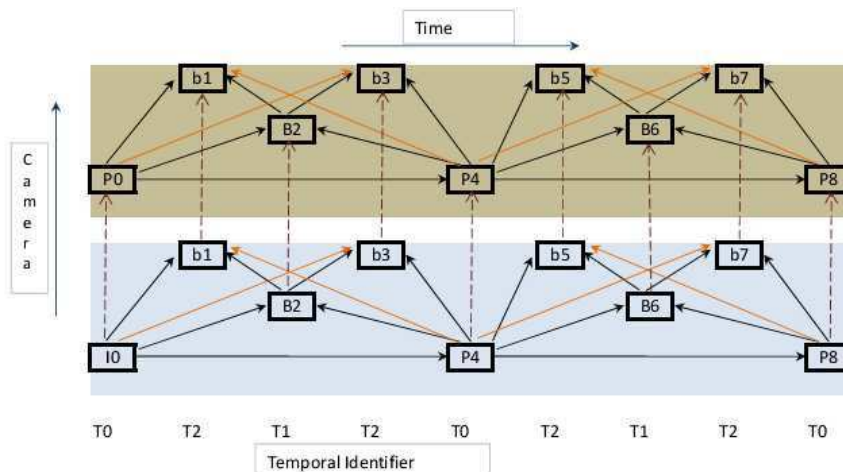
One of the major factors in the improvement of compression efficiency in case of disposable view components based hierarchical coding is the cascading of QP from one layer to another. Unfortunately, this is not possible for the prediction structure shown in Fig.5.2, as even the highest layer view components are used for reference. But the disposable view components concept could be combined with the temporal id based method and the resultant prediction structure is shown in Fig.5.3.



**Figure 5.2:** Temporal id based hierarchical coding with three temporal layers

### 5.1.3. Proposed prediction structure

The combined prediction structure shown in Fig.5.3 will provide reduced compression efficiency when compared to the disposable view components based hierarchical coding as shown in Fig.5.1, as the number of temporal predictors is less. For example, consider P8 in the left view, the number of temporal predictors is ‘2’ in Fig.5.1, whereas it is single in Fig.5.3. On the other hand, the combined prediction structure (Fig.5.3) provides full temporal scalability using the temporal\_id when compared to the disposable view components based hierarchical coding, which facilitates the production of only one sub sequence by discarding the disposable ‘b’ view components which are present at the highest temporal layer.



**Figure 5.3:** Hierarchical coding with temporal id and disposable b frames combined



#### 5.1.4. Simulation setup and results

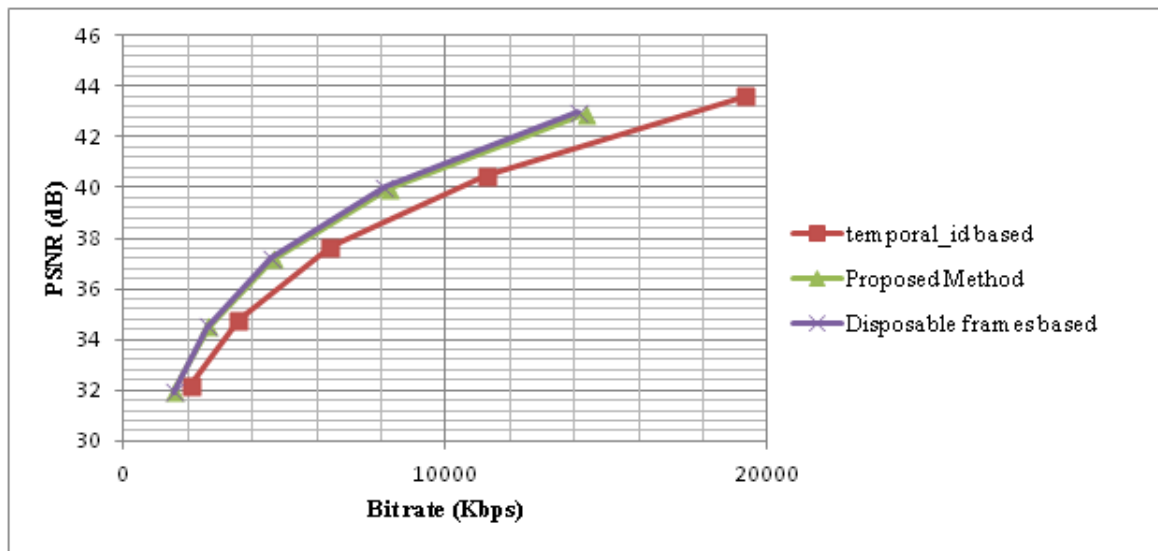
Simulations have been setup to evaluate the rate-distortion performance of all the three methods (disposable view components based, temporal identifier based and proposed method) discussed above. For this simulation, H.264/AVC reference software (JM) has been used. Since JM doesn't support temporal scalability, the reference software has been enhanced to include the support for the same. I.e. for the evaluation of disposable view components method, the JM code is used as-is and for the evaluation of the other two methods, the enhanced source code (which includes the support for temporal id based temporal scalability) is used. Frames from cameras zero and one, for the Akko and Kayo, Crowd, Flamenco, Objects and race MVC sequences of VGA resolution[58] as well as the Balloons and Kendo sequences [59] of XGA resolutions are used for the evaluation. Table III show the coding tools used in JM and Bjøntegaard measurements [62] are used to compare coding performance between the coding structures (with disposable view components based method as reference) and the results for IBBP and IPPP cases are shown in Table IV. Negative value depicts that "disposable view components based" method outperforms the temporal id based/proposed method by such margin. It could be seen that the maximum reduction in DPSNR (Delta PSNR) of the proposed method is 0.34 dB, where as in case of temporal id based method it is around 2 dB. The R–D performance curve for the crowd sequence is shown in Fig.5.4.

**Table III:** Coding tools used in JM for the simulation of temporal scalability

Coding tools used for the simulation	
Software	JM 18.2 (H.264 / MVC)
Profile	Stereo High Profile
Resolution	640 x 480 (VGA), 1280x768 (XGA)
Frame rate	30
Number of B frames	3
Number of reference frames	3
ME search range	32
QP values used	20 - 36 (QP for B/b is +2)
Software	JM 18.2 (H.264 / MVC)

**Table IV :** Bjontegaard R-D performance numbers, DPSNR is in units of dB and DRate is in units of kbps

Test Video clip information		Temporal id based (IBBP)		Proposed method (IBBP)		Temporal id based (IPPP)		Proposed method (IPPP)	
Resolution	Sequence	DSNR	DRate	DSNR	DRate	DSNR	DRate	DSNR	DRate
VGA 640x480 30 Hz	Akko& Kayo	-2	45.06	-0.05	0.87	-1.89	40.22	-0.05	1.08
	Crowd	-1.21	27.3	-0.09	1.83	-1.51	35.92	-0.09	1.83
	Flamenco	-1.02	24.83	-0.07	1.48	-1.26	31.75	-0.07	1.48
	Objects	-1.11	24.26	-0.34	6.72	-1.42	33.22	-0.34	6.72
	Race	-0.96	24.75	-0.06	1.39	-1.41	39.66	-0.06	1.39
XGA 1024x768 30 Hz	Balloons	-0.81	21.15	-0.02	0.53	-0.48	13.18	-0.04	0.87
	Kendo	-0.68	18.11	-0.01	0.31	-0.37	10.13	-0.01	0.19



**Figure 5.4:** R-D performance for Crowd sequence

## 5.2. Improved temporal scalability for H.264 MVC

The temporal scalability restriction in terms of the prediction structure could be stated as,

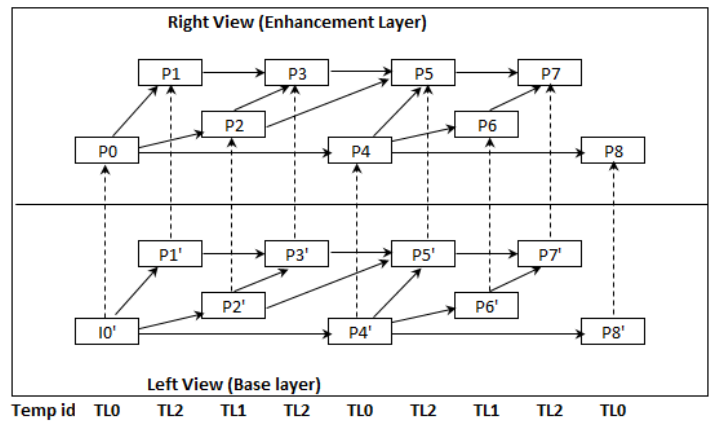
$$tempid(ref) \leq tempid(curr)$$

Where  $tempid(ref)$  is the temporal identifier of the reference slice and  $tempid(curr)$  is the temporal identifier of the current slice. This condition ensures that pictures present at a particular temporal level TLN could be removed at the media gateway without affecting the decoding process of TL (N-1). It is clear from the above condition that, the reference pictures in the constructed list should have the temporal index which is less than or equal to the temporal index of the current slice. But unfortunately, in all the variants of H.264 (SVC, MVC and MVC + Depth), temporal identifier information is used only in the sub bit extraction process, but not used in the reference picture list construction process.

### 5.2.1. Reference picture list construction process

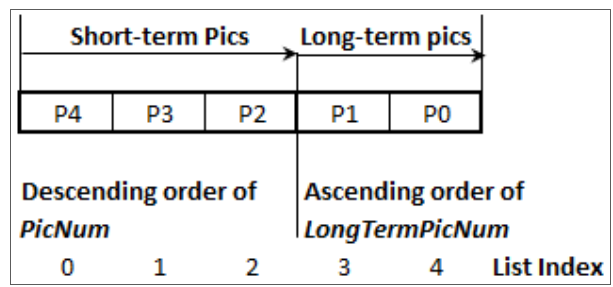
The concept of short-term and long-term reference pictures have been introduced in H.263++ and it has been shown that long-term reference frames combined with multiple reference frames improves the coding efficiency [46][63]. This idea has been extended to H.264/AVC and H.264 MVC as well and there a number of algorithms proposed for fast selection of the best reference picture [64], selection of subset of reference pictures to improve the rate-distortion performance [65]. An example prediction structure is shown in Fig.5.5, which provides three levels of temporal scalability ("TL0", "TL1" and "TL2"). Since there are multiple reference pictures, default reference picture lists (which may hold the subset of multiple reference pictures) will be constructed based on a pre-defined process and if the encoder wish to re-organize the default constructed list, the same could be achieved by using reference picture list re-ordering (RPLR) commands, sent at every slice header. Short-term reference pictures are identified by using picture number (PicNum) and long term reference pictures are identified by using long term picture number (LongTermPicNum). The default list construction process in H.264/AVC works as follows: First the short-term reference pictures are arranged in the descending order of PicNum and long term reference pictures are arranged in the ascending order of LongTermPicNum. As a next step, re-arranged short-term pictures are placed first in the list followed by the re-arranged long-term pictures, such that the short-term reference pictures will have lower list indices compared to long-term reference pictures. After the default list is constructed, additional entries beyond "number of active reference pictures" for the slice will be discarded. The default list construction process ensures that there is at least one short-term reference picture and at least one

"available for reference" picture is present at the end of construction process in the list. Fig.5.6 depicts the structure of the default reference picture list.



**Figure 5.5:** Prediction structure for a three layer temporal scheme using H.264 MVC. Straight lines indicate temporal prediction and dashed lines indicate inter-view prediction.

For the prediction structure of right view pictures shown in Fig.5.5 (for both base and enhancement layer), there are two ways to realize temporal scalability. In the first method, pictures at the lowest temporal level (“TL0”) will be coded as a long term picture and subsequent frames could refer to this long term picture till the next “TL0” level picture arrives, which will replace the earlier “TL0” picture, by using MMCO commands. The other way is to code all the pictures as short-term reference pictures and by using RPLR commands, the default reference picture list could be modified at every slice level to suit the requirement. It has been shown that using long-term reference pictures for temporal scalability results in degradation in compression efficiency [66] compared to using short term reference pictures, thus the latter method is considered in the following sections.



**Figure 5.6:** Default arrangement of reference picture list considering current slice as P5 for the enhancement layer prediction structure shown in Fig.5.5. It is assumed that the pictures P1 and P0 are coded as long term pictures.

The default constructed list before encoding picture P5 (with respect to enhancement layer prediction structure shown in Fig.5.5), which is present in temporal level “TL1” is shown in Fig.5.7. Whereas the required state of the

list should only have the frames which are present in “TL1” or “TL0” level, as shown in Fig.5.8. In order to achieve this, RPLR commands will be sent in all the slices of P5, which consume a good amount of bits.

	P4	P3	P2	P1	P0	P5'
<b>Temporal Level</b>	TL0	TL2	TL1	TL2	TL0	TL1
<b>List Index</b>	0	1	2	3	4	5

**Figure 5.7:** Default state of reference picture list before the coding of picture P5 with respect to right view prediction structure shown in Fig.5.5. All reference pictures are coded as short-term pictures and three temporal layer scheme is employed.

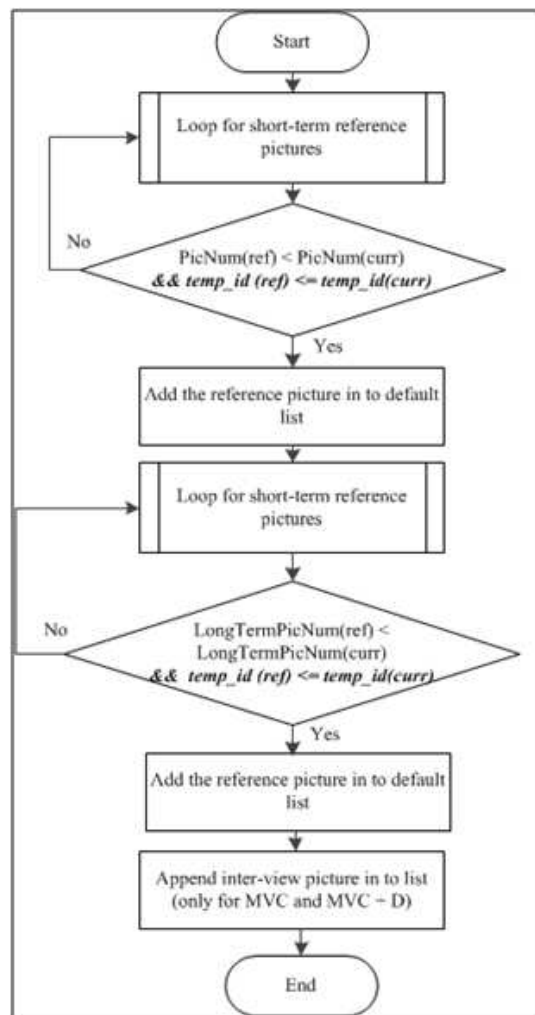
	P4	P2	P0	P5'	P1	P3
<b>Temporal Level</b>	TL0	TL1	TL0	TL1	TL2	TL2
<b>List Index</b>	0	1	2	3	4	5

**Figure 5.8:** Modified reference picture list before coding of P5 to suit temporal scalability requirement for the right-view prediction structure shown in Fig.5.5. Temporal level of P5 is TL1 and the reference pictures should be present at TL0 and TL1. P5' is the inter-view picture which is also coded at temporal level TL1.

### 5.2.2. Modified list construction process

As mentioned in Sec.5.2.1, the reference picture list reordering needs to be performed in order to support temporal scalability. This reference picture list reordering w.r.t temporal scalability requirements could be avoided by using the modified list construction process as outlined below. The modified reference picture list construction process works as follows: a short-term picture will be added into the reference picture list if the PicNum of the reference slice is less than the PicNum of the current slice and “temporal identifier of the reference slice is less than or equal to the temporal identifier of the current slice”. Similarly a long-term picture will be added into the reference picture list if LongTermPicNum of the reference slice is less than the LongTermPicNum of the current slice and “temporal identifier of the reference slice is less than or equal to the temporal identifier of the current slice”. The flow chart which depicts the existing "default list construction process" with the proposed modifications mentioned in bold italicized text is shown in Fig.5.9. The resultant list will contain reference pictures with a temporal identifier which is less than or equal to the current picture's temporal identifier, thereby avoiding the need of sending the RPLR commands to suit temporal scalability requirement. The proposed modifications are possible for the enhancement layer slice, as temporal identifier information is available for H.264 MVC (in NAL unit header MVC extension). It could be noted that the temporal identifier information is not available for the base layer and hence the proposed method is applicable only for enhancement layers.

Consider the same scenario as depicted in Fig.5.7, for frame P5 (which is present at Temporal level TL1), when the default list construction process with the proposed modifications is used, the constructed list will be exactly same as in Fig.5.8, thereby avoiding the need for sending the RPLR commands. There are a number of advantages of the proposed method; first it ensures that there is no need to spend bits for RPLR to realize required list arrangement to satisfy temporal scalability. Second, other valid pictures could be accommodated in the list, as the pictures beyond “active number of reference pictures” will be discarded in the existing method. For example in case of H.264 MVC, when the modified method is used, the “additional” inter-view pictures could be placed in the default constructed list, which improves R-D performance. Also in the case of slice header loss, the modified method ensures that the state of both encoder and decoder default constructed lists are same.



**Figure 5.9:** Flow chart for the default list construction process for H.264 variants and the italicized text depicts the proposed modifications [P1]

### 5.2.3. Simulation setup and results

For this simulation H.264/AVC reference software JM [60] which also supports encoding of stereoscopic video using H.264 MVC is used with the prediction structure as shown in Fig.5.5. Test clips of various resolutions have been used to evaluate the R-D performance and results are illustrated in Table V. The results show that the maximum bit rate savings achieved for the case of full HD (1080p) is around 310 Kbps for the depicted stereoscopic case, with PSNR reduction of only 0.02 dB. The maximum number of views supported by H.264 MVC is 1024, which includes 1023 enhancement views and a base view. Thus, in theory around 300 Mbps bit rate savings could be achieved in the case of full HD for the three temporal layer scheme. Only one list is considered in the simulation (as the prediction structure consists of predictive slices only), but there will be two lists when bi-predictive slices are used and the savings in bit rate will be more. This clearly indicates that, the results show the lower bound of achievable savings in bit rate. This will be useful in scenarios where the number of views is high, especially in the case of free view point TV.

The delta bit rate and delta PSNR values are calculated by using the following formulae,

$$\text{delta bit rate (kbps)} = \text{Default list's bit rate(kbps)} - \text{Modified list's bit rate (kbps)}$$

$$\text{delta PSNR (dB)} = \text{Default list's PSNR (dB)} - \text{Modified list's PSNR (dB)}$$

The values of average delta bit rate and average delta PSNR are calculated by using the following formulae,

$$\text{Average delta bit rate (kbps)} = \sum \text{delta bit rate (kbps)} \div \text{number of bit streams}$$

$$\text{Average delta PSNR (dB)} = \sum \text{delta PSNR (dB)} \div \text{number of streams}$$

**Table V:** R-D performance comparison between existing and modified H.264 MVC schemes for a three temporal layer and stereoscopic video case

YUV	QP	Default list		Modified list		Delta Bit rate (Kbps)	Delta PSNR (dB)	Avg. Delta bit rate (Kbps)	Avg. Delta PSNR (dB)
		Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)				
Crowd VGA, 30 Hz	20	22153.9	43.42	22068.3	43.415	85.61	-0.006	72.71	0.00
	24	13039.35	40.24	12958.6	40.239	80.75	-0.005		
	28	7404.44	37.40	7333.2	37.397	71.29	-0.003		
	32	4083.78	34.48	4030.6	34.478	53.2	-0.005		
Flamenco VGA, 30 Hz	20	9524.09	45.48	9500.2	45.467	23.91	-0.012	26.67	-0.01
	24	5642.22	43.27	5613.5	43.256	28.73	-0.013		
	28	3363.32	41.00	3332.7	40.99	30.66	-0.014		
	32	1977.61	38.27	1954.3	38.261	23.36	-0.012		
Race VGA, 30 Hz	20	10376.71	43.88	10207.7	43.87	169.04	-0.007	131.65	-0.01
	24	5780.43	41.17	5634.5	41.168	145.93	-0.005		
	28	3130.04	38.58	3007.6	38.573	122.44	-0.007		
	32	1623.2	35.89	1534.0	35.875	89.17	-0.015		
Balloons XGA, 30 Hz	20	11371.34	45.25	11068.7	45.243	302.64	-0.01	179.79	0.00
	24	5785.78	43.57	5584.6	43.567	201.17	-0.005		
	28	3405.69	41.74	3272.5	41.736	133.18	-0.003		
	32	2169.41	39.31	2087.2	39.309	82.18	0		
Kendo XGA, 30 Hz	20	10106.07	46.28	9632.2	46.264	473.88	-0.018	308.36	-0.02
	24	5540.33	44.68	5186.8	44.661	353.58	-0.019		
	28	3309.53	42.89	3061.7	42.869	247.84	-0.018		
	32	2046.82	40.57	1888.7	40.557	158.14	-0.011		
<b>Max average delta bit rate and delta PSNR</b>								<b>308.36</b>	<b>-0.02</b>



## **Chapter 6**

### **Spatial scalability**

This chapter introduces the various coding tools for providing the spatial scalability option, signaling of these tools to the decoder and benchmarking data (rate-distortion performance) with respect to simulcast option. Since the up-sampling filter plays an important role in the compression efficiency of spatial scalability a separate simulation has been set up to evaluate a couple of different up-sampling filters. Also the combined spatiotemporal scalability option is presented in the chapter.

#### **6.1. Inter-layer prediction**

Inter-layer prediction refers to the scenario where the reference layer pixels are used as the prediction signal in the enhancement layer. In previous standards, such as H.262/MPEG-2, H.263 and MPEG-4 Visual there are two ways by which the reference layer pixels are used,

- Up-sampled reference layer pixels are used as the final prediction signal in the enhancement layer
- Reference layer pixels will be up-sampled and averaged with the enhancement layer's temporal signal and the resultant pixels will be used as prediction signal in the enhancement layer

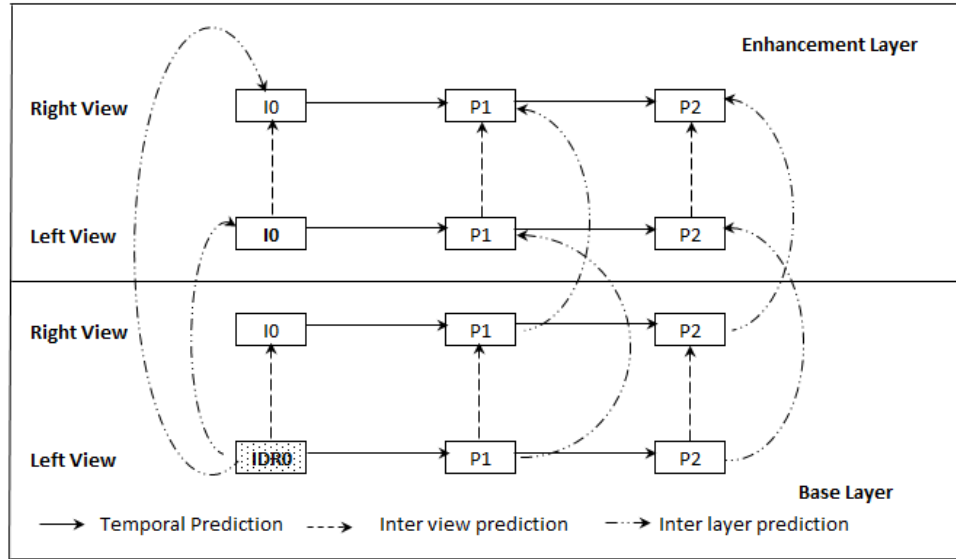
The usage of up-sampled reference layer pixels as prediction signal in the enhancement layer will not be the best approximation of original video content always. For example, in segments of the video sequence with low motion and high spatial detail, temporal signal (contents from the adjacent macroblocks / contents from the co-located macroblock from the previous frames) will provide a better rate-distortion performance than the reference layer pixels. But in these scenarios, the enhancement layer shares a number of similar aspects with the reference layer such as intra prediction modes, motion vectors, reference picture list information and this redundancy in coding the header bits could be exploited. In addition to this, there will be similarities between the reference layer residual pixels and enhancement layer residual pixels as they both represent similar if not same video content. In order to exploit the above-mentioned redundancies to improve the coding efficiency, the following inter-layer coding tools based on H.264 scalable video coding [67] have been introduced in the proposed scalable multiview video coding compression scheme,

- Inter-layer intra prediction
- Inter-layer motion prediction
- Inter-layer residual prediction

The SMVC encoder will evaluate both the intra- and inter-layer prediction tools and choose the one which could provide a better representation of the content to be encoded using minimum number of bits. Inter-layer prediction has been designed in such a way that the enhancement layer will have a higher spatial identifier 'D' when compared to the reference layer. Also in order to ensure parallel encoding/decoding of views in the layers, only the corresponding view of the reference layer is used for prediction in the enhancement layer. The design supports spatial scalability with arbitrary resolution ratios, with enhancement layer has a picture width and/or height more than the width and/or height of the reference layer as in the case of H.264 SVC [68- 69]. Though the design allows arbitrary ratios, for the sake of simplicity, the description of the inter-layer prediction techniques have been limited to the case of dyadic spatial scalability with enhancement layer's width / height is twice the width / height of the reference layer.

### **6.1.1. Inter-layer intra prediction**

Inter-layer intra prediction refers to the scenario, where the reference layer reconstructed pixels could be used as a prediction signal at the enhancement layers. As the reconstruction process involves parsing, decoding and motion compensation, it consumes a good amount of computing power at the decoder and hence should be applied only to the selected views/layers. Though the process is computationally intensive, it serves as a synchronization point for the decoders, in the case of packet loss. The base view of the base layer (which is backward compatible with H.264/AVC) has been considered as the synchronization point and only this view component will be reconstructed and all the views in the enhancement layers (both base and enhancement views) will use these reconstructed pixels as the prediction signal. Moreover, since inter layer intra prediction is applicable only for intra MBs, the reconstruction process at the synchronization point will be invoked only when the corresponding MB is coded as intra, as this avoids the scenario where the neighboring inter Macro blocks(MBs) needs to be reconstructed, which require multiple loop decoding. Fig.6.1 depicts various prediction schemes for stereoscopic case with a base layer and one enhancement layer and the synchronization point is marked in dotted box. Since only the reconstructed intra MB pixels are available, constrained intra prediction will be enabled in all the view components of all the layers and in the case of spatial scalability the predicted pixels will be up-sampled before using them as the predictors in the enhancement layers. It could be noted that the performance of this tool depends heavily on the up-sampling filter used, as the prediction error could be minimized by using an appropriate filter.



**Figure 6.1:** Various prediction schemes employed by the SMVC method. Dotted lines represent temporal prediction, dashed lines represent inter-view prediction and curved dash-dot lines represent inter-layer prediction.

### 6.1.2. Inter-layer motion prediction

As the scene content remains same between the layers, there will be high correlation among the layers and inter layer motion prediction enables the re-use of reference layer motion information at the enhancement layers. The motion vectors which belong to the reference layers will be parsed at the decoding end and in the case of spatial scalability, the same will be up-scaled before using them at the enhancement layer. More specifically,

$$MV_{curr} = MV_{ref} \times \text{spatial ratio}$$

$MV_{curr}$  is the motion vector at the current layer and  $MV_{ref}$  is the motion vector of the reference layer (parsed from the reference layer bit-stream). Spatial ratio is the ratio between the current and enhancement layers and is given by,

$$\text{Spatial ratio} = \text{Enhancement layer width (height)} \div \text{Base layer width (height)}$$

There are two options to use the derived motion information from the reference layer; it could be used as the final motion vector (termed as base mode option) or it could be used as the predicted motion vector (motion prediction option) in the enhancement layer. When the base mode option is chosen as the best, the motion vectors, reference indices related information will not be coded in the bitstream and the same will be inferred from the reference layer. On the other hand, when the motion prediction option is chosen by the encoder, difference between the current layer's motion vector and reference layer's motion vector will be coded in the bitstream. Also to ensure parallel decoding, the motion information from the corresponding view of the reference layer will be

used as the predictor in the enhancement layer. For example, in the prediction structure depicted in Fig.6.1, the motion information of the left view of base layer is used as the predictor for the left view of enhancement layer and the base layer's right view motion information is used as a predictor in the enhancement layer's right view.

### **6.1.3. Inter-layer residual prediction**

The inter-layer residual prediction tool exploits the redundancies exist in the residual domain and because the inter-layer intra prediction already exploited redundancies in the pixel domain, this tool will not be applied for intra MBs. The residual pixels of the reference layers will be parsed, inverse quantized, inverse transformed and used as prediction signal at the enhancement layers. At the enhancement layers, the encoder will subtract the reference layer residual from the current layer's residuals and only the difference between the residuals will be subject to transform coding process. In the case of spatial scalability, the reference layer residuals will be up-sampled before using them as the prediction signal. At the decoding end, the reference layer residuals (decoded and up-sampled in the case of spatial scalability) will be added to the current layer residuals before adding the resultant with the predicted pixels. The up-sampling filter (bi-linear poly phase FIR filter), used in the H.264 SVC specification [70] with respect to inter-layer residual prediction is used in SMVC case also and similar to the case of inter-layer motion prediction, in order to ensure parallel decoding, the residual pixels from the corresponding view of the reference layer will be used as the predictor in the enhancement layer.

### **6.1.4. Disparity compensated prediction**

H.264 MVC specification already supports disparity compensated prediction to a certain extent for the inter slices in the form of inter-view prediction [11], but not for the intra slices. In other words, for the inter slice, when the disparity compensated prediction is better than the temporal prediction, the encoder will simply re-arrange the reference picture list, as the initial state of the decoded picture buffer will have temporal view components first followed by the view components from other views. Consider the scenario in Fig.6.1, as mentioned in Sec.6.1.1, only the left view of the base layer (synchronization point) will be reconstructed and for the right view of the enhancement layer (coded as an intra-slice), the reconstructed pixels could not be used as-is and needs to be disparity compensated. For this, the disparity offset vectors in the base layer are calculated by treating the right view component as current frame and left view component as reference frame and full search motion estimation is carried out for every MB. At the enhancement layer (for the right view components), the disparity vectors will be added with the macro block positions and the resultant co-ordinates will be used to fetch the reference area in the base view of the base layer and the same will be used as the prediction signal. Again, for the case of spatial scalability, the disparity offsets will be up-scaled (as explained in Sec.6.1.2) before using them as predictors.

### 6.1.5. Single loop decoding

Multiple loop decoding was one of the major reasons for the failure of scalable video coding of MPEG-2 and MPEG-4[68], as it requires motion compensation to be carried out at the reference layers in addition to the current layer, which increases the decoder complexity. On the other hand, single loop decoding ensures that any tool at the enhancement layer will have minimal impact at the decoder side, as it requires the decoding of layer ‘N’ as well as the parsing, motion compensation of the (N-1)<sup>th</sup> layer. In the proposed scalable multi view coding framework, it is ensured that decoding of only the appropriate layer (the layer for which the decoder is configured) is needed and minimal parsing of the other reference layers are needed. The only exception to the concept of single loop decoding in the proposed SMVC scheme is the inter layer intra prediction tool, which involves decoding of the reference layer information such that the decoded pixels will be used as prediction signal at the enhancement layer. But as mentioned in the previous sections, the reconstruction is limited to the base view of the base layer only and also the dependency between the views across the layers have been kept in such a way that all the prediction information (motion vector, reference indices, disparity offsets, residual pixels) of View ‘N’ in enhancement layer depends only on View ‘N’ of the reference layers, which ensures parallel parsing/decoding.

### 6.1.6. Signaling mechanisms

At the NAL level, the presence of spatial / coarse grain scalability information could be signaled by using the reserved bit (renamed as S-bit, for the purpose of SMVC) which exists in the 3-byte extended NAL header. For the base layer NAL packets, the S-bit will be coded as zero and for the enhancement layer NAL packets, the S-bit will be coded as one. The SMVC compatible decoder will use this S – bit, along with the picture width and height information in the sequence parameter set (SPS) of both base and enhancement layers to determine the change in spatial resolution between the layers (spatial scalability / coarse grain bit rate scalability), whereas the MVC compatible decoder will ignore this bit. At the SPS level, new profile information (scalable\_multi\_view\_high\_profile) is coded for the enhancement layers, whereas the base layer could be coded using stereo high / multi view high profile, with the base view of base layer coded at high profile. At the slice level the following parameters are coded for the enhancement layer slices.

- **adaptive\_base\_mode\_flag:** Indicates whether the “base mode” flag (Inter-layer intra prediction in case of intra MBs and up-scaled motion vectors of reference layer as final motion vectors in the enhancement layer for the case of Inter MBs) information is signaled in the MB header or slice header. If the value of the flag is zero, then this indicates that the base mode flag is signaled at the slice level and is applicable for all the MBs of the slice. Else, this information is signaled at the MB level.

- **Slice\_level\_base\_mode\_flag:** Present when the `adaptive_base_mode_flag` is zero and possible values for the flag will be zero (base mode option is disabled for all the MBs of the slice) or one (base mode option is enabled for all the MBs of the slice).
- **adaptive\_motion\_pred\_flag:** Indicates whether the “motion prediction” flag (motion vectors of reference layer as predicted motion vectors in the enhancement layer for the case of Inter MBs) information is signaled in the MB header or slice header. If the value of the flag is zero, then this indicates that the “motion prediction” flag is signaled at the slice level and is applicable for all the MBs of the slice. Else, this information is signaled at the MB level.
- **Slice\_level\_motion\_pred\_flag:** Present when the `adaptive_motion_pred_flag` is zero and the possible values for the flag will be one (motion prediction option is enabled for all the MBs of the slice) or zero.
- **adaptive\_residual\_pred\_flag:** Indicates whether the “residual prediction” flag (motion vectors of reference layer as predicted motion vectors in the enhancement layer for the case of Inter MBs) information is signaled in the MB header or slice header. If the value of the flag is zero, then this indicates that the “residual prediction” flag is signaled at the slice level and is applicable for all the MBs of the slice. Else, this information is signaled at the MB level.
- **Slice\_level\_residual\_pred\_flag:** Present when the `adaptive_residual_pred_flag` is zero and the possible values for the flag will be one (residual prediction option is enabled for all the MBs of the slice) or zero.

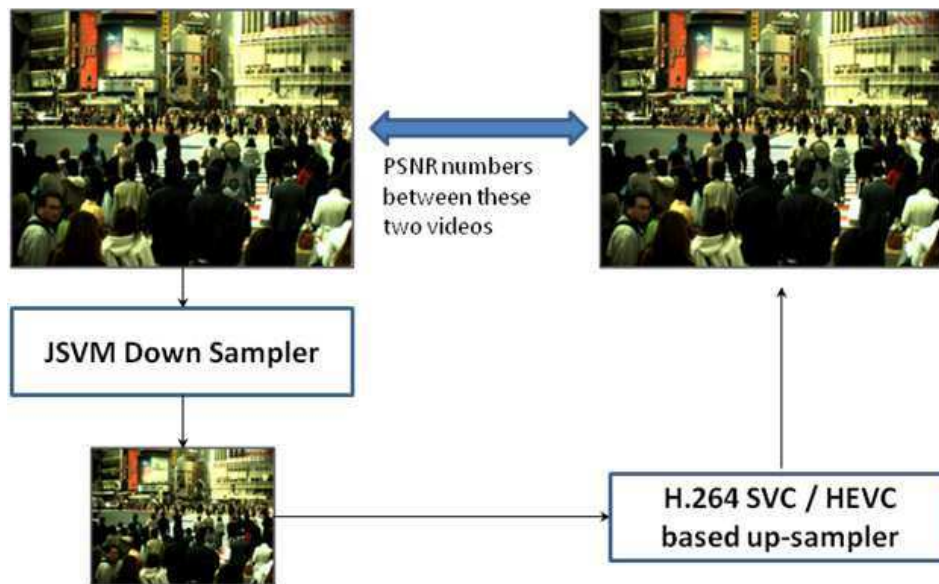
At the MB level the following parameters are coded for the enhancement layer macro blocks.

- **MB\_level\_base\_mode\_flag:** Present when the `adaptive_base_mode_flag` is enabled in the slice header and the possible values for the flag will be zero (base mode option is disabled for the current MB) or one.
- **MB\_level\_motion\_pred\_flag:** Present when the `adaptive_motion_pred_flag` is enabled in the slice header and the possible values for the flag will be zero (motion prediction is disabled for the current MB) or one.
- **MB\_level\_residual\_pred\_flag:** Present when the `adaptive_residual_pred_flag` is enabled in the slice header and the possible values for the flag will be zero (residual prediction option is disabled for the current MB) or one.
- **Disparity\_x\_offset:** Encoded for the intra MBs corresponding to the enhancement views of the enhancement layers, in the units of full pixel. Specifies the offset in the horizontal direction, using which, predicted pixels could be fetched from the base view of the base layer.
- **Disparity\_y\_offset:** Similar to `Disparity_x_offset`, but specifies the offset in the vertical direction.

## 6.2. Simulation setup and results

### 6.2.1. Comparison between H.264 SVC and HEVC Up-sampling filters

As mentioned in the previous sections, for the spatial scalability case, the compression performance of the inter-layer intra prediction tool depends heavily on the up-sampling filters. In this simulation setup, the compression performance and speed of the conversion process are evaluated for the different filters. There are two choices considered in the simulation namely the H.264 SVC up-sampling filter [31] and the High Efficiency Video Coding (HEVC) up-sampling filter [71]. In order to evaluate the PSNR performance of the two filters, the left and right views of the test videos are dyadic down-sampled by using the down-sampling filter provided by the JSVM reference software [72]. The down-sampling filter is a 12-tap poly-phase FIR filter with filter kernel of  $\{1, 2, -5, -9, 18, 57, 57, 18, -9, -5, 2, 1\}$  (for the phase value of 8, corresponding to dyadic case) [73] with appropriate rounding and division and the resultant videos are up-sampled by using H.264 SVC and HEVC filters. Finally, the PSNR numbers are measured between the original and the up-sampled video as explained in Fig.6.2.



**Figure 6.2:** PSNR and speed evaluation process between H.264 SVC and HEVC based up-sampling filters. Process shows only the left view of the crowd sequence and the same will be repeated for right view as well and the average between the two values will be computed.

Only stereoscopic case is considered for this simulation and frames from cameras 0 and 1 of various MVC test sequences [58] are used for the evaluation. Table VI depicts the PSNR and conversion performance for various test sequences, where the numbers represents the average values of left and right views. It could be seen that the HEVC based filter provides approximately 3 dB gain but consumes around 3x – 4x more time when compared to

the H.264 SVC based filter. Since the up-sampling is needed only for the intra MBs corresponding to base view of the base layer, the additional processing time will not be an overhead and the HEVC based filter is used in the SMVC scheme. Hence in all the subsequent simulations, HEVC based up-sampling filter is used.

**Table VI:** PSNR and speed of conversion numbers for H.264 SVC and HEVC based filters. Performance measured on an Intel core i3 laptop clocked at 2.4 GHz with 4GB RAM

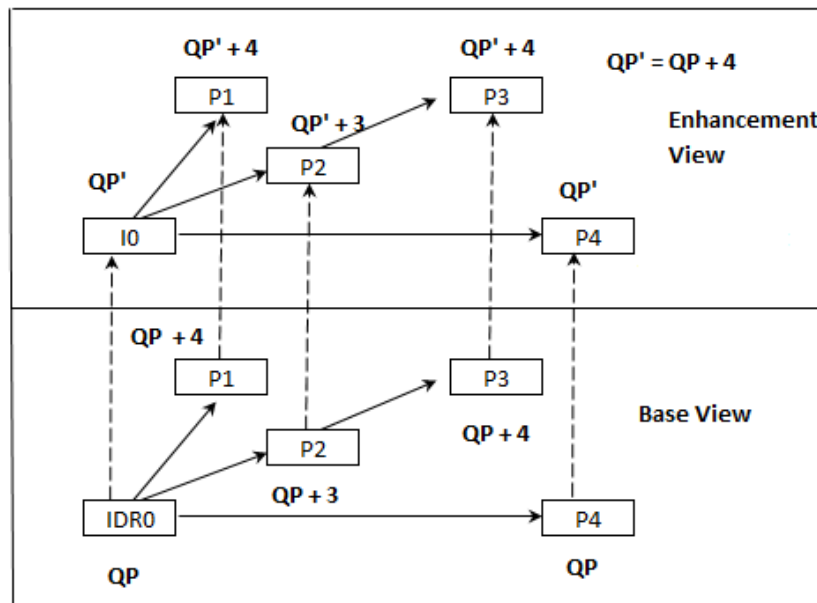
Sequence Name	PSNR performance (dB)			Conversion performance (ms)		
	H.264 SVC filter	HEVC filter	Difference	H.264 SVC filter	HEVC filter	Difference
Akko & Kayo	31.89	35.13	3.25	0.315	1.245	-0.93
Crowd	25.08	28.07	2.99	0.34	1.22	-0.88
Flamenco	32.96	36.05	3.08	0.675	1.185	-0.51
Objects	22.79	25.65	2.85	0.34	1.02	-0.68
Race	27.53	30.44	2.91	0.655	1.185	-0.53
<b>Average PSNR performance difference (dB)</b> <b>3.01645</b>				<b>Average conversion performance difference (ms)</b> <b>-0.706</b>		

### 6.2.2. Evaluation for spatial scalability

In this simulation setup, the R-D performance of the proposed SMVC method is compared with the simulcast method for the case of spatial scalability. The SMVC scheme has been implemented using JM 18.2 [60] and stereoscopic video (with left and right views) has been considered for the evaluation with one base layer and an enhancement layer. In other words, two layers have been considered with each layer contains two views. Camera frames from 0 and 1, for the Akko & Kayo, Crowd, Flamenco, Objects and race MVC sequences of VGA resolution are used for the evaluation. The base layer is encoded at a QP which is higher than that of enhancement layer QP (to simulate the scenario where the enhancement layer bit rate will be higher than the base layer bit rate) and a hierarchical prediction structure with QP cascading scheme as shown in Fig.6.3 is employed. The input videos are down-sampled by using a linear separable filter with kernel values of  $\{-8, 0, 24, 48, 48, 24, 0, -8\}$  normalized by 128 with rounding [68]. Only dyadic ratio (where the height and width of the base layer sequence is half of enhancement layer sequence) is considered. The R-D performance numbers for both the proposed SMVC method and simulcast method are mentioned in Table VII. The bitstream extractor tool provided by the



JMVC reference software [55] has been modified to extract the NAL packets of base layers and Table VIII illustrates the R-D performance numbers for the base layer (H.264 MVC compatible) and base view of base layer (H.264 AVC compatible). The R-D curves for the flamenco and race sequences are shown in Fig.6.4 and Fig.6.5 respectively. From the table as well as from the R-D graphs, it could be seen that the proposed SMVC method provides around 10% reduction in bit rate with a max PSNR reduction of around 0.07 dB. In the simulations, only predictive slices are used and when bi-predictive slices are used, savings in bit rate will be more. Also, it could be noted that the R-D performance could be improved by using intelligent, input video dependent filters and the mentioned R-D results provide only a lower bound on achievable performance.



**Figure 6.3:** Hierarchical prediction structure and the QP cascading scheme used in the simulation setup

The gain in bit rate and average gain in bit rate values mentioned in the below table are computed using,

$$Gain\ in\ bit\ rate\ (\%) = (Simulcast\ method\ bit\ rate\ (Kbps) - Proposed\ method\ bit\ rate\ (Kbps)) \div Simulcast\ method\ bit\ rate\ (Kbps) * 100$$

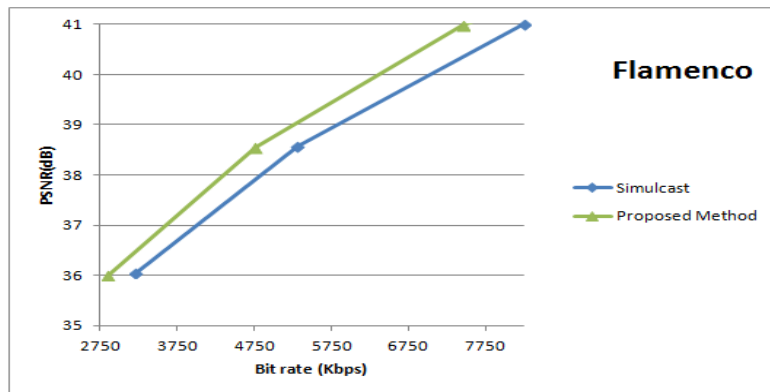
$$Average\ gain\ in\ bit\ rate\ (\%) = \sum Gain\ in\ bit\ rate\ (\%) \div number\ of\ streams$$

**Table VII:** R-D performance figures of simulcast and proposed SMVC methods for spatial scalability

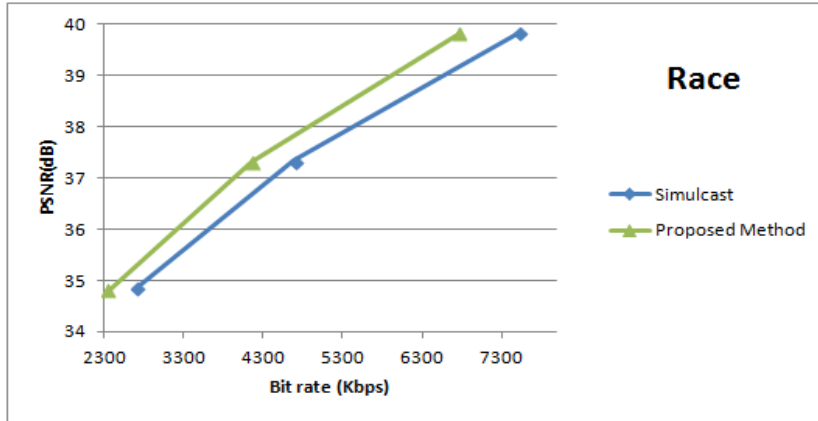
Sequence name	QP	Simulcast		Proposed Method		Gain in bit rate %	Avg gain in bit rate %	Delta PSNR	Avg Delta PSNR
		Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)				
Akko & Kayo	BL:26, EL:24	8974.17	40.76	8190.02	40.73	8.74	10.0	-0.03	-0.04
	BL:30, EL:28	5470.25	38.34	4923.55	38.31	9.99		-0.04	
	BL:34, EL:32	3175.9	35.86	2815.26	35.8	11.36		-0.06	
Crowd	BL:26, EL:24	13105.62	39.31	12238.47	39.28	6.62	7.7	-0.02	-0.03
	BL:30, EL:28	8201.52	36.71	7578.14	36.69	7.6		-0.02	
	BL:34, EL:32	4831.31	34.16	4399.9	34.13	8.93		-0.04	
Flamenco	BL:26, EL:24	8244.02	41	7450.64	40.98	9.62	10.3	-0.02	-0.03
	BL:30, EL:28	5306.5	38.58	4758.02	38.55	10.34		-0.03	
	BL:34, EL:32	3211.41	36.06	2853.04	36.01	11.16		-0.04	
Objects	BL:26, EL:24	4196.78	39.49	3895.28	39.44	7.18	8.2	-0.06	-0.07
	BL:30, EL:28	2649.64	36.65	2431.91	36.58	8.22		-0.07	
	BL:34, EL:32	1627.42	33.7	1477.06	33.61	9.24		-0.09	
Race	BL:26, EL:24	7523.05	39.84	6761.34	39.81	10.13	11.5	-0.02	-0.04
	BL:30, EL:28	4690.33	37.34	4157.65	37.3	11.36		-0.03	
	BL:34, EL:32	2707.89	34.87	2350.52	34.81	13.2		-0.06	

**Table VIII:** R-D performance figures for base layer and base view of base layer for spatial scalability

Sequence name	QP	Base layer		Base view of base layer	
		Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)
Akko & Kayo	BL:26, EL:24	1828.12	39.23	869.66	39.29
	BL:30, EL:28	1112.66	36.59	523.86	36.64
	BL:34, EL:32	676.57	33.97	316.18	34.05
Crowd	BL:26, EL:24	2667.31	37.99	1681.58	36.63
	BL:30, EL:28	1663.86	35.18	1067.44	33.64
	BL:34, EL:32	1036.42	32.49	665.8	30.84
Flamenco	BL:26, EL:24	1767.04	39.65	878.25	39.50
	BL:30, EL:28	1091.78	36.92	541.78	36.77
	BL:34, EL:32	662.65	34.28	327.19	34.13
Objects	BL:26, EL:24	805.54	38.81	375.36	38.79
	BL:30, EL:28	526.58	35.60	247.33	35.54
	BL:34, EL:32	354.51	32.48	167.06	32.40
Race	BL:26, EL:24	1301.14	38.97	659.95	38.88
	BL:30, EL:28	817.22	36.27	415.08	36.17
	BL:34, EL:32	525.97	33.68	267.93	33.58



**Figure 6.4:** R-D performance curve for Flamenco sequence for the spatial scalability case



**Figure 6.5:** R-D performance curve for Race sequence for the spatial scalability case

The compression ratio and encoder processing speed measurements for all the above-mentioned simulation videos are shown in Table IX. The performance numbers were measured on an Intel core-i5 laptop, with 4 GB RAM and clocked at 3.0 GHz. The multi-view video numbers (depicted in the last column of Table IX) are projected numbers for three views with estimated complexity increase of 50%, when compared to stereo video encoding. The reason for the “estimation” of encoding speed for multi view videos is that the JM reference software used in the simulations support only stereo video encoding and the other available option of using the JMVC reference software results in frequent application failures during the encoding process. The justification behind the estimate are,

- There will not be an increase in the encoding speed of the reference view (view 0 and view 1) frames, as higher view frames will not be used as reference in the motion estimation process.
- There is an additional frame (corresponding to the third view) that needs to be encoded and during the encoding of this frame, the number of reference frames will be same as that of reference layer frames even though there will be an increase in the number of inter-view reference frames.
- In other words “compute intensive” modules like motion estimation will take same amount of time w.r.t other views’ encoding as the additional inter-view frames will compete with the temporal predictor for a place in the decoded picture buffer. Also, the number of inter-layer predictors will remain same when compared to other reference layers.

**Table IX:** Compression ratio and encoder processing speed for Spatial Scalability

Sequence name	QP	Bit rate (original video) (Bytes per sec)	Bit rate (compressed video) (Bytes per sec)	Compression ratio	Encoder frame rate (frames per second) for stereo video	Encoder frame rate (frames per second) for multi-view video
Akko & Kayo	BL:26, EL:24	34560000	1023752.5	34	10.2	5.1
	BL:30, EL:28	34560000	615443.75	56	10.27	5.14
	BL:34, EL:32	34560000	351907.5	98	10.57	5.29
Crowd	BL:26, EL:24	34560000	1529808.7	23	9.63	4.82
	BL:30, EL:28	34560000	947267.5	36	10.26	5.13
	BL:34, EL:32	34560000	549987.5	63	10.18	5.09
Flamenco	BL:26, EL:24	34560000	931330	37	9.93	4.97
	BL:30, EL:28	34560000	594752.5	58	9.49	4.75
	BL:34, EL:32	34560000	356630	97	9.35	4.68
Objects	BL:26, EL:24	34560000	486910	71	9.78	4.89
	BL:30, EL:28	34560000	303988.75	114	10.43	5.22
	BL:34, EL:32	34560000	184632.5	187	10.62	5.31
Race	BL:26, EL:24	34560000	845167.5	41	9.18	4.59
	BL:30, EL:28	34560000	519706.25	66	9.02	4.51
	BL:34, EL:32	34560000	293815	118	9.74	4.87

The bit rate of the original video in the units of bytes per second is calculated as follows,

$$\begin{aligned}
 EL/BL \text{ Bit rate (bytes per sec)} \\
 &= \text{frame width} \times \text{frame height} \times YUV \text{ space factor} \times \text{frame rate} \\
 &\times \text{number of views}
 \end{aligned}$$

$$\text{Original video Bit rate (bytes per sec)} = BL \text{ Bit rate (bytes per sec)} + EL \text{ Bit rate (bytes per sec)}$$

The bit rates of the compressed video are derived using the below formula,

$$\text{Bit rate ( bytes per sec)} = \text{Bit rate (kilo bits per sec)} \times 1000 \div 8$$

$$\text{Compression ratio} = \text{Original video bit rate (bytes per sec)} \div \text{Compressed bit rate (bytes per sec)}$$

The frame rates depict the processing speed and is excluding the FILE Input / Output timings. It could be noted that there is no multi-threading implementation and application is using single core. The frame rates could be increased by processing individual views of the enhancement layer in separate thread as all the tools of the proposed SMVC framework w.r.t view  $V_N$  depends only on the corresponding view in the reference layer. Also the frame rates could be increased by using “early termination” algorithms which prevents the “exhaustive” evaluation of inter and intra layer prediction mechanisms before finalizing the best prediction option.

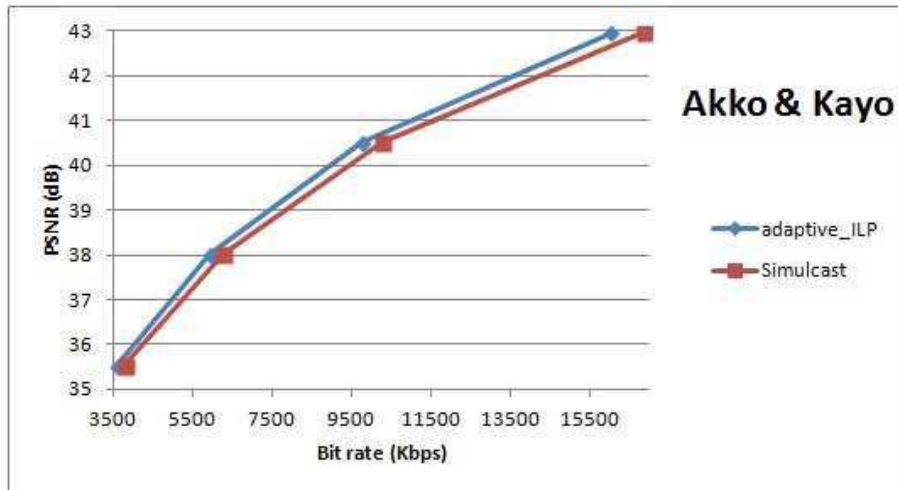
The compression ratio is less for scenarios where the quantization is on the lower side (Eg: BL: 26, EL: 24), this is expected as the intention was to preserve as much of source video content as possible. Also the enhancement layer QP is chosen in such a way that it is less than that of base layer as the bit rate of the enhancement layer is expected to reduce at the enhancement layers.

### **6.2.3. Spatio-temporal scalability**

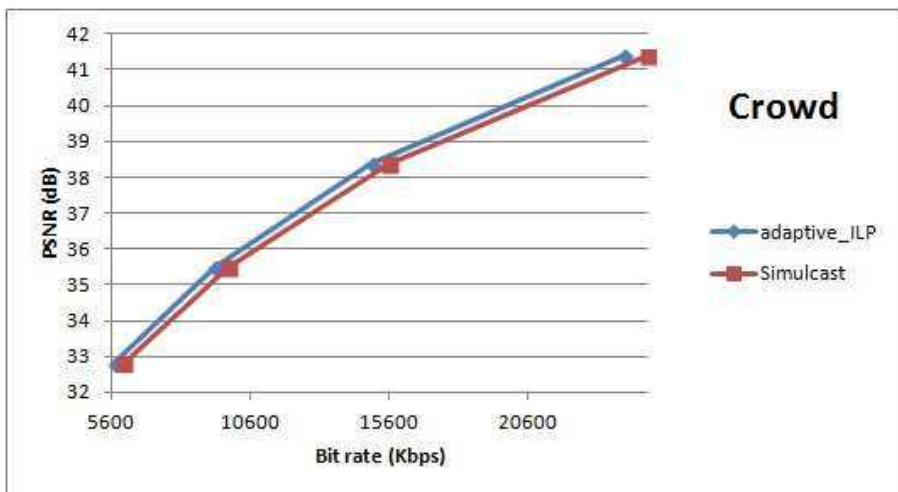
In this simulation setup, the R-D performance of the proposed SMVC method is compared with the simulcast method for the case of spatio-temporal scalability. The prediction structure employed in the previous sections has been used for this simulation as well and the down-sampling filter as employed in Sec.6.2.1 is used. The enhancement layer, (higher spatial resolution video) is encoded at a frame rate of 'N' and the base layer (low resolution video) is encoded at a frame rate of 'N/2' and the input video will be down sampled before feeding the same as the input to base layer encoder. The simulation results are shown in Table X. Also, the rate-distortion curves are depicted in Fig.6.6 – Fig.6.9 for various test sequences such as Akko & Kayo, Crowd, Flamenco and Race respectively.

**Table X:** R-D performance results for Spatio-temporal scalability

Sequence	QP	Proposed method		Simulcast method		Delta		Percentage bit rate Gain (%)
		Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)	
Crowd	BL:22, EL:20	24101.5	41.3795	24996.7	41.38	-895.22	-0.0005	4.66
	BL:26, EL:24	14956.3	38.368	15640.1	38.3675	-683.87	0.0005	5.78
	BL:30, EL:28	9305.37	35.508	9809.15	35.509	-503.78	-0.001	6.87
	BL:34, EL:32	5651.3	32.8025	6004.88	32.8155	-353.58	-0.013	8.02
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-0.0035</b>	<b>6.33</b>
Flamenco	BL:22, EL:20	16103.5	43.3035	16854.3	43.315	-750.77	-0.0115	5.97
	BL:26, EL:24	10467.5	40.8055	11046.0	40.811	-578.57	-0.0055	7.04
	BL:30, EL:28	6867.16	38.2375	7293.24	38.243	-426.08	-0.0055	7.83
	BL:34, EL:32	4300.11	35.6305	4583.78	35.64	-283.67	-0.0095	8.36
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-0.008</b>	<b>7.30</b>
Akko& Kayo	BL:22, EL:20	16002.5	42.9395	16835.1	42.967	-832.58	-0.0275	6.56
	BL:26, EL:24	9741.24	40.509	10258.5	40.527	-517.28	-0.018	6.67
	BL:30, EL:28	5903.44	38.005	6249.98	38.0205	-346.54	-0.0155	7.33
	BL:34, EL:32	3538.67	35.503	3747.64	35.5315	-208.97	-0.0285	7.50
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-0.02238</b>	<b>7.01</b>
Race	BL:22, EL:20	23724.7	41.9045	25288.9	41.91	-1564.13	-0.0055	7.77
	BL:26, EL:24	14939.5	39.2075	16277.2	39.1945	-1337.64	0.013	10.37
	BL:30, EL:28	9548.99	36.5635	10561.2	36.5575	-1012.3	0.006	12.15
	BL:34, EL:32	5824.81	34.0265	6561.8	34.0205	-736.99	0.006	14.47
<b>Average delta PSNR and percentage bit rate gain</b>							<b>0.00487</b>	<b>11.19</b>

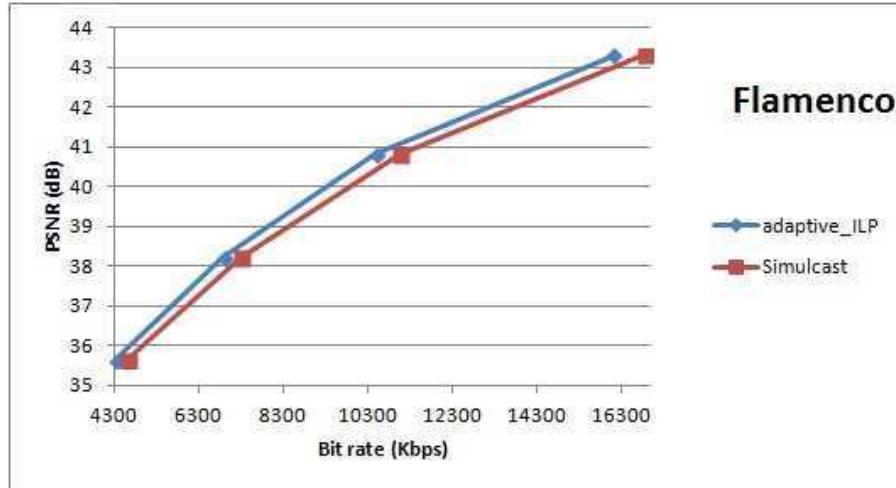


**Figure 6.6:** R-D curves for Akko & Kayo sequence: Spatio-temporal scalability

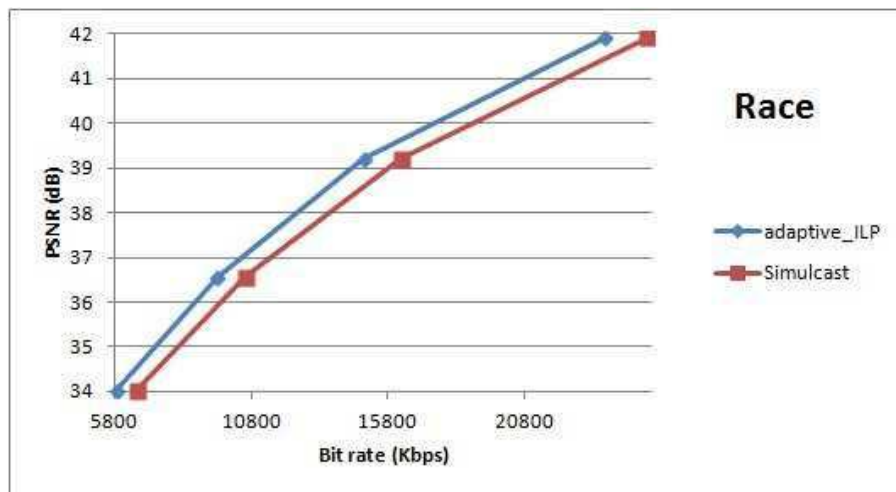


**Figure 6.7:** R-D curves for Crowd sequence: Spatio-temporal scalability





**Figure 6.8:** R-D curves for Flamenco sequence: Spatio-temporal scalability



**Figure 6.9:** R-D curves for Race sequence: Spatio-temporal scalability

From Table X and the R-D performance curves, it could be noted that the SMVC method provides around 8% reduction in bit rate with average PSNR reduction of only 0.01 dB. Similar to Sec.6.2.2, in the simulations, only predictive slices are used and when bi-predictive slices are used, savings in bit rate will be more. Also, it could be noted that the R-D performance could be improved by using intelligent, ‘content-aware’ filters and the mentioned R-D results provide only a lower bound on achievable performance.

The compression ratio and encoder speed timings (in terms of frames per second) are shown in Table XI. The formulae used to compute the bit rate of the original, compressed videos and compression ratio are same as that of spatial scalability scenario mentioned in Sec.6.2.2. The performance numbers were measured on an Intel core-i5 laptop, with 4 GB RAM and clocked at 3.0 GHz. The multi-view video numbers (depicted in the last column of Table XI) are projected numbers for three views with estimated complexity increase of 50%. The justification behind the estimate mentioned in Sec.6.2.2 holds good for this section as well. Again, the frame rates depict the processing speed and is excluding the FILE Input / Output timings. Similar to the spatial scalability scenario, it could be noted that there is no multi-threading implementation and application is using single core. The frame rates could be increased by processing individual views of the enhancement layer in separate thread as well as with the usage of “early termination” algorithms.

**Table XI:** Compression ratio and encoder processing speed for Spatio-temporal scalability

Sequence name	QP	Bit rate (original video) (Bytes per sec)	Bit rate (compressed video) (Bytes per sec)	Compression ratio	Encoder frame rate (frames per second) for stereo video	Encoder frame rate (frames per second) for multi-view video
Crowd	BL:22, EL:20	34560000	3012687.5	11	13.39	6.70
	BL:26, EL:24	34560000	1869537.5	18	14.45	7.22
	BL:30, EL:28	34560000	1163171.25	30	15.39	7.70
	BL:34, EL:32	34560000	706412.5	49	15.27	7.64
Flamenco	BL:22, EL:20	34560000	2012937.5	17	14.28	7.14
	BL:26, EL:24	34560000	1308437.5	26	14.90	7.45
	BL:30, EL:28	34560000	858395	40	14.24	7.12
	BL:34, EL:32	34560000	537513.75	64	14.03	7.01
Akko & Kayo	BL:22, EL:20	34560000	2000312.5	17	15.12	7.56
	BL:26, EL:24	34560000	1217655	28	15.3	7.65
	BL:30, EL:28	34560000	737930	47	15.41	7.70
	BL:34, EL:32	34560000	442333.75	78	15.86	7.93
Race	BL:22, EL:20	34560000	2965587.5	12	13.04	6.52
	BL:26, EL:24	34560000	1867437.5	19	13.77	6.89
	BL:30, EL:28	34560000	1193623.75	29	13.53	6.77
	BL:34, EL:32	34560000	728101.25	47	14.61	7.31

## Chapter 7

### Bit rate scalability

This chapter introduces the following types of bit rate scalability options,

- Coarse grain bit rate scalability
- Medium grain bit rate scalability

The former option could be considered as the special case of spatial scalability with no change in spatial resolution across the layers and is useful when the bit rate variation in the network is high. The latter option could be accomplished by using the temporal scalability feature and is useful when there are minimal variations in the bit rate. It is worth mentioning that the medium grain bit rate scalability option could be used to generate subsequences. Rate-distortion performance have been benchmarked for both the above-mentioned options as well as for the combined CGS-temporal scalability. Part of the chapter summarizes the [J2] publication.

#### 7.1. Coarse grain (CGS) bit rate scalability

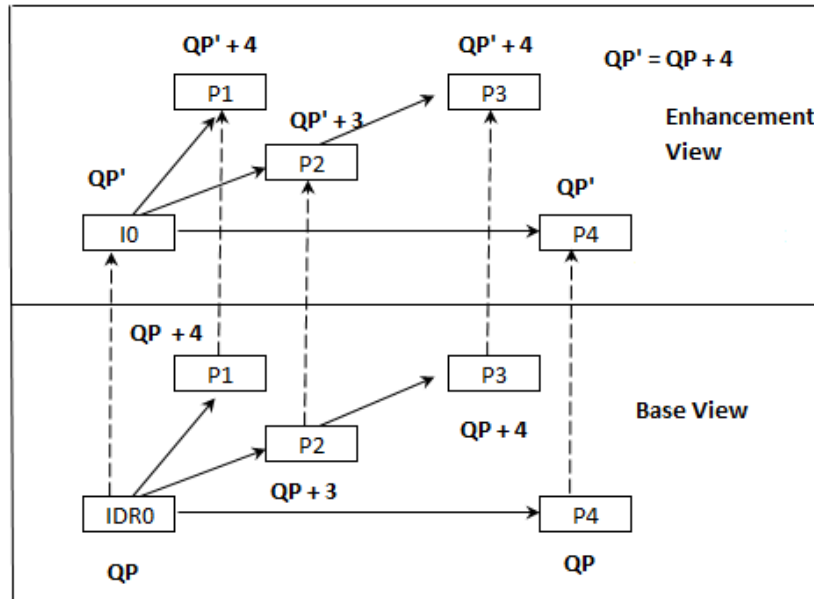
As mentioned, this option could be considered as the special case of spatial scalability with no change in spatial resolution across the layers. The purpose of this simulation is to bench mark the R-D performance of the proposed SMVC scheme with the simulcast method. The SMVC scheme has been implemented using JM 18.2 [60] and stereoscopic video (with left and right views) has been considered for the evaluation with one base layer and an enhancement layer. In other words, two layers have been considered with each layer contains two views. Camera frames from 0 and 1, for the Akko & Kayo, Crowd, Flamenco, Objects and race MVC sequences of VGA resolution are used for the evaluation. The base layer is encoded at a QP which is higher than that of enhancement layer QP' (to simulate the scenario where the enhancement layer bit rate will be higher than the base layer bit rate) and a hierarchical prediction structure with QP cascading scheme as shown in Fig.7.1 is employed. The R-D performance numbers for both the proposed SMVC method and simulcast method are mentioned in Table XII. Bitstream extractor tool provided by the JMVC reference software [55] has been modified to extract the NAL packets of base layers and Table XIII illustrates the R-D performance numbers for the base layer (H.264 MVC compatible) and base view of base layer (H.264 AVC compatible). The R-D curves for the flamenco and race sequences are shown in Fig.7.2 and Fig.7.3 respectively. From the table as well as from the R-D graphs, it could be seen that the proposed SMVC method provides around 20% reduction in average bit rate with a max PSNR reduction of around 0.23 dB.

**Table XII:** R-D performance figures of simulcast and proposed SMVC methods for coarse grain scalability

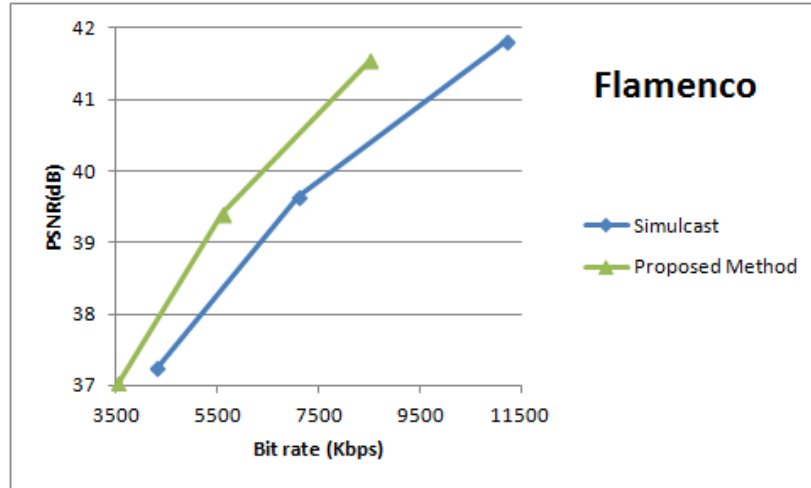
Sequence name	QP	Simulcast		Proposed Method		Gain in Bit rate %	Avg gain in bit rate %	Delta PSNR	Avg Delta PSNR
		Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)				
Akko & Kayo	BL:26, EL:24	12551.95	41.69	9981.3	41.46	20.48	18.7	-0.23	-0.20
	BL:30, EL:28	7481.58	39.47	6063.71	39.29	18.95		-0.18	
	BL:34, EL:32	4441.94	37.11	3696.16	36.93	16.79		-0.18	
Crowd	BL:26, EL:24	22602.22	38.38	17511.55	38.12	22.52	19.5	-0.26	-0.21
	BL:30, EL:28	13958.73	35.91	11157.22	35.72	20.07		-0.19	
	BL:34, EL:32	8373.09	33.54	7020.93	33.37	16.15		-0.17	
Flamenco	BL:26, EL:24	11202.39	41.81	8496.88	41.55	24.15	20.9	-0.26	-0.23
	BL:30, EL:28	7087.55	39.64	5597.69	39.41	21.02		-0.23	
	BL:34, EL:32	4309.33	37.25	3549.51	37.05	17.63		-0.20	
Objects	BL:26, EL:24	5865.52	39.57	4667.05	39.36	20.43	18.7	-0.21	-0.22
	BL:30, EL:28	3624.03	36.97	2917.37	36.73	19.50		-0.23	
	BL:34, EL:32	2302.25	34.30	1929.12	34.09	16.21		-0.22	
Race	BL:26, EL:24	11486.44	40.08	9056.78	39.88	21.15	19.7	-0.20	-0.18
	BL:30, EL:28	7023.29	37.75	5648.54	37.58	19.57		-0.17	
	BL:34, EL:32	4226.86	35.48	3441.06	35.31	18.59		-0.16	

**Table XIII:** R-D performance figures for base layer and base view of base layer for coarse grain scalability

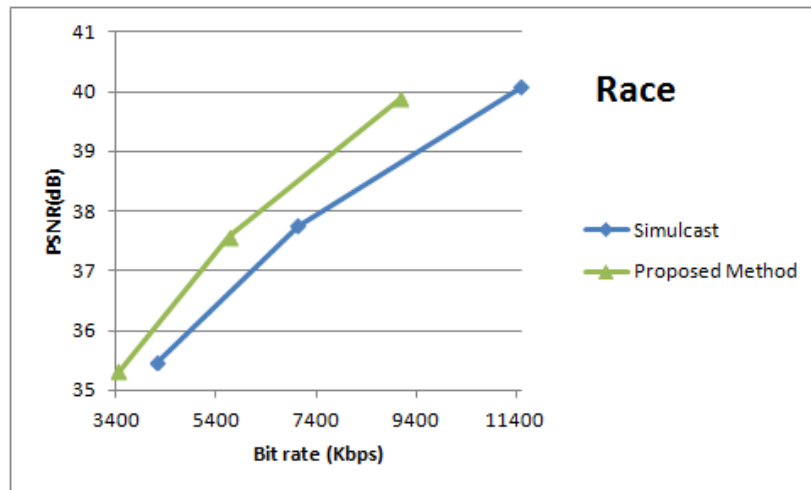
Sequence name	QP	Base layer		Base view of base layer	
		Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)
Akko & Kayo	BL:26, EL:24	5260.92	41.11	2527.69	41.11
	BL:30, EL:28	3008.4	38.87	1431.22	38.89
	BL:34, EL:32	1852.31	36.51	872.19	36.54
Crowd	BL:26, EL:24	9624.52	37.72	4988.2	37.50
	BL:30, EL:28	5695.22	35.29	2952.08	35.04
	BL:34, EL:32	3569.24	32.99	1846.91	32.73
Flamenco	BL:26, EL:24	4722.63	41.27	2423.37	40.83
	BL:30, EL:28	2871.98	39.04	1461.77	38.59
	BL:34, EL:32	1761.07	36.68	888.78	36.28
Objects	BL:26, EL:24	2466.09	38.97	1181.75	38.86
	BL:30, EL:28	1497.91	36.24	715.7	36.05
	BL:34, EL:32	1024.96	33.69	490.04	33.48
Race	BL:26, EL:24	4964.09	39.47	2561.42	39.31
	BL:30, EL:28	2902.59	37.14	1494.22	36.96
	BL:34, EL:32	1864.6	34.94	954.83	34.78



**Figure 7.1:** Hierarchical prediction structure and the QP cascading scheme used in the simulation setup



**Figure 7.2:** R-D performance curve for Flamenco sequence for the coarse grain scalability case



**Figure 7.3:** R-D performance curve for Race sequence for the coarse grain scalability case

The compression ratio and encoder speed timings (in terms of frames per second) are shown in Table XIV. The formulae used to compute the bit rate of the original, compressed videos and compression ratio are same as that of spatial scalability scenario mentioned in Sec.6.2.2. The performance numbers were measured on an Intel core-i5 laptop, with 4 GB RAM and clocked at 3.0 GHz.

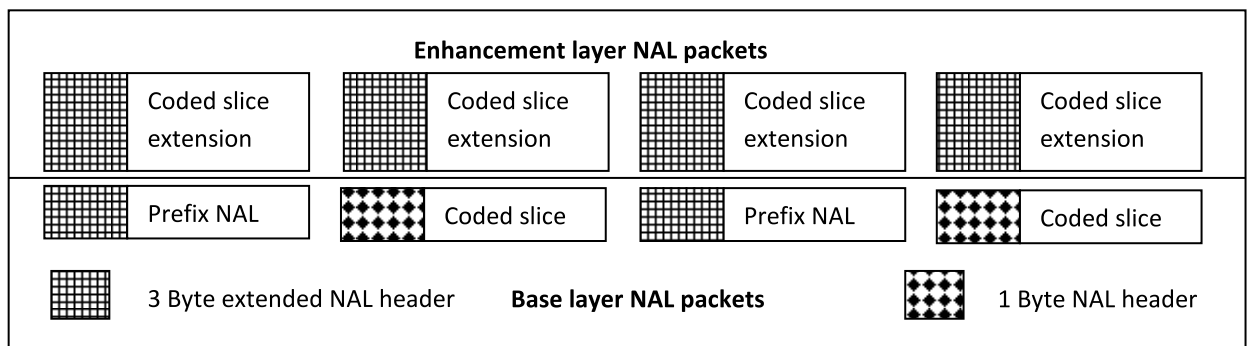
**Table XIV:** Compression ratio and encoder processing speed for coarse grain scalability

Sequence name	QP	Bit rate (original video) (Bytes per sec)	Bit rate (compressed video) (Bytes per sec)	Compression ratio	Encoder frame rate (frames per second) for stereo video	Encoder frame rate (frames per second) for multi-view video
Akko & Kayo	BL:26, EL:24	55296000	1247662.5	44	6.79	3.40
	BL:30, EL:28	55296000	757963.75	73	6.95	3.48
	BL:34, EL:32	55296000	462020	120	7.13	3.57
Crowd	BL:26, EL:24	55296000	2188943.75	25	6.77	3.39
	BL:30, EL:28	55296000	1394652.5	40	6.92	3.46
	BL:34, EL:32	55296000	877616.25	63	7.06	3.53
Flamenco	BL:26, EL:24	55296000	1062110	52	6.98	3.49
	BL:30, EL:28	55296000	699711.25	79	7.07	3.54
	BL:34, EL:32	55296000	443688.75	125	7.2	3.60
Objects	BL:26, EL:24	55296000	583381.25	95	6.84	3.42
	BL:30, EL:28	55296000	364671.25	152	6.96	3.48
	BL:34, EL:32	55296000	241140	229	6.69	3.35
Race	BL:26, EL:24	55296000	1132097.5	49	6.22	3.11
	BL:30, EL:28	55296000	706067.5	78	6.43	3.22
	BL:34, EL:32	55296000	430132.5	129	6.63	3.32

The bit rate corresponding of the original video is more when compared to the spatial scalability scenario, as mentioned in Table and Table. The reason behind this increase is the fact that in case of spatial scalability, the base layer is encoded at reduced resolution when compared to the original video, where-as in the case of CGS, the resolutions are same in both base layer as well as enhancement layers. The multi-view video numbers (depicted in the last column of Table XIV) are projected numbers for three views with estimated complexity increase of 50%. The justification behind the estimate mentioned in Sec.6.2.2 holds good for this section as well. Again, the frame rates depict the processing speed and is excluding the FILE Input / Output timings. Similar to the spatial scalability and spatio-temporal scalability scenarios, it could be noted that there is no multi-threading implementation and application is using single core. The frame rates could be increased by processing individual views in separate thread as all the tools of the proposed SMVC framework in the current view depends only on the corresponding view in the reference layer. Also the usage of “early termination” algorithms will help the encoder in arriving at the best prediction option without performing an exhaustive evaluation of all the inter-layer and intra-layer prediction options.

## 7.2. Medium grain bit rate scalability

Consider a scenario where there are a number of 3D smart televisions connected to a MANE, which is receiving the 3D video from a production studio. In order to cope with the variations in the available bandwidth, the packets which belongs to the higher temporal layers (in both base view and enhancement view) needs to be dropped in the MANE and the process to identify the packets' temporal identifier should be lightweight as deep inspection of the packets will lead to significant end-to-end video delay. The problem here is that the temporal identifier is present only for the enhancement layer packets and the same are not available for the base layer packets. In order to support backward compatibility with the existing 2D receivers, in H.264 multi view coding specification, base layer NAL packets will be preceded by a prefix NAL as depicted in Fig.7.4, which could be used for the `temporal_id` interpretation process.

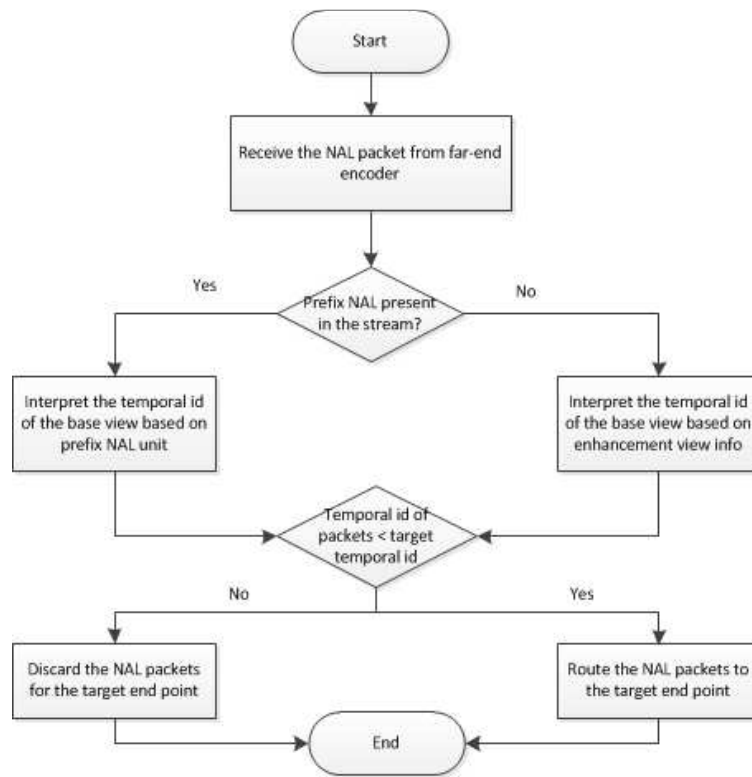


**Figure 7.4:** Packet arrangement in H.264 MVC

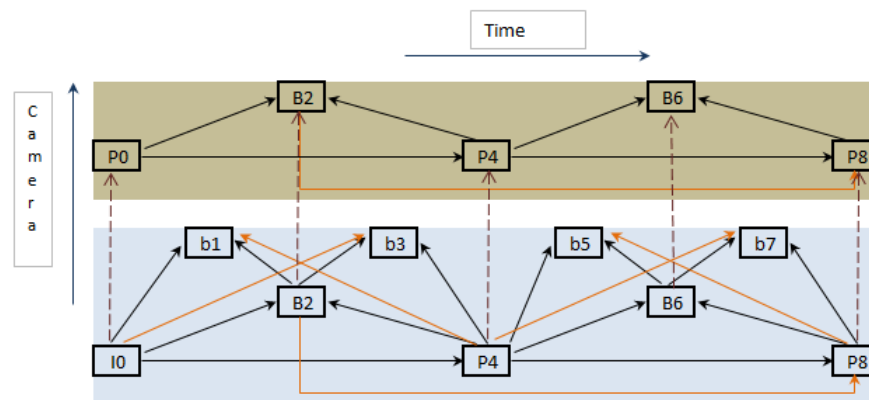
Similar to enhancement layer NAL packets, prefix NAL contains 3-byte extended NAL header that contains information related to the succeeding base layer NAL packet. One of the fields in the extended NAL is the `temporal_id` information. Hence MANE can “interpret” the temporal identifier of the base layer NAL packets using the preceding prefix NAL information. However, the prefix NAL is an optional component and encoders can choose not to send the prefix NAL and under these circumstances, the temporal identifier of the base layer packets could be “interpreted” using the temporal identifier of the enhancement layer NAL packets, as the definition of an access unit in H.264 MVC specification includes the view components captured at the same time from all the views. Hence the value of the temporal identifier would be the same for all the view components which contains same frame number (i.e. captured at the same time). The flow chart depicting the interpretation process outlined above is shown in Fig.7.5. When the available bandwidth drops further down, the enhancement view components could be sent at a lower frame rate when compared to the base view component as for inter-view prediction, the enhancement view components depends on the base view and the other way around is not



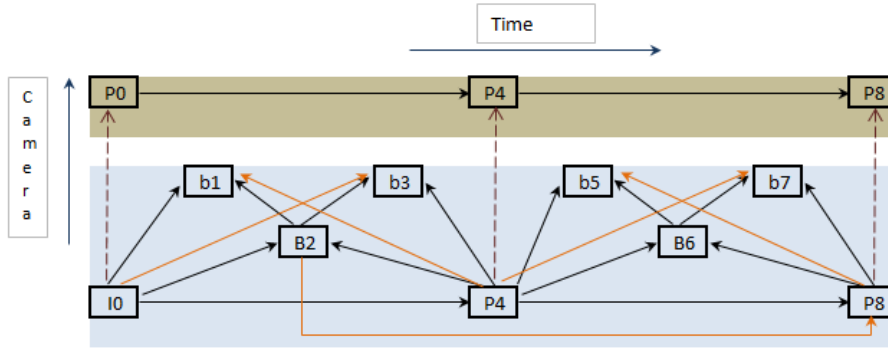
true. The sample resultant prediction structures are shown in Fig.7.6 and Fig.7.7, where the enhancement views are coded at  $\frac{1}{2}^{\text{nd}}$  and  $\frac{1}{4}^{\text{th}}$  of frame rate when compared to the base view.



**Figure 7.5:** Interpretation process for temporal identifier of base view



**Figure 7.6:** Enhancement view at  $\frac{1}{2}^{\text{nd}}$  of base view frame rate



**Figure 7.7:** Enhancement view at 1/4th of base view frame rate

The simulation has been setup to show that a number of subsequences could be extracted from the encoded stream with minimal computations. The bitstream extractor software provided by JMVC reference software [55] is used, since the extractor software doesn't support temporal identifier based extraction, the same has been added (as depicted in Fig.7.5). Encoded streams from first simulation have been used as input and it could be seen that five different sub sequences could be extracted from a single encoded bitstream with 3 temporal layers. The results are depicted in Table XV and Table XVI for the IBBP and IPPP case respectively. It could be seen that the enhancement views could be extracted at a lower frame rate when compared to the base view. This is possible, because the enhancement views depend on the base view for inter-view prediction and the required reference base view component for an enhancement view component will always be available. Also, the average extraction time for a view component is as low as 0.38 ms (measured on an Intel core-i5 laptop, with 4 GB RAM and clocked at 3.0 GHz) as deep inspection of the packets at MANE will not be possible.

**Table XV:** Percentage bit rates and average frame extraction time for IBBP case. Frame rates are with respect to original encoded frame rate. 1/2 means 50% with respect to original frame rate.

Sequence name	Bit rate percentages at lower frame rates (%)					Avg. extraction time /frame(ms)
	v0 = 1 v1 = 1/2	v0 = 1 v1 = 1/4	v0 = 1/2 v1 = 1/2	v0 = 1/2 v1 = 1/4	v0 = 1/4 v1 = 1/4	
Akko& kayo	93.06	82.45	86.89	76.28	66.62	0.25
Crowd	90.67	79.89	79.84	69.07	57.06	0.28
Flamenco	90.28	79.39	78.35	67.46	54.88	0.26
Objects	93.02	86.01	86.52	79.52	73.51	0.25
Balloons	90.22	80.93	79.51	70.22	60.36	0.33

**Table XVI:** Percentage bit rates and average frame extraction time for IPPP case. Frame rates are with respect to original encoded frame rate.  $\frac{1}{2}$  means 50% with respect to original frame rate.

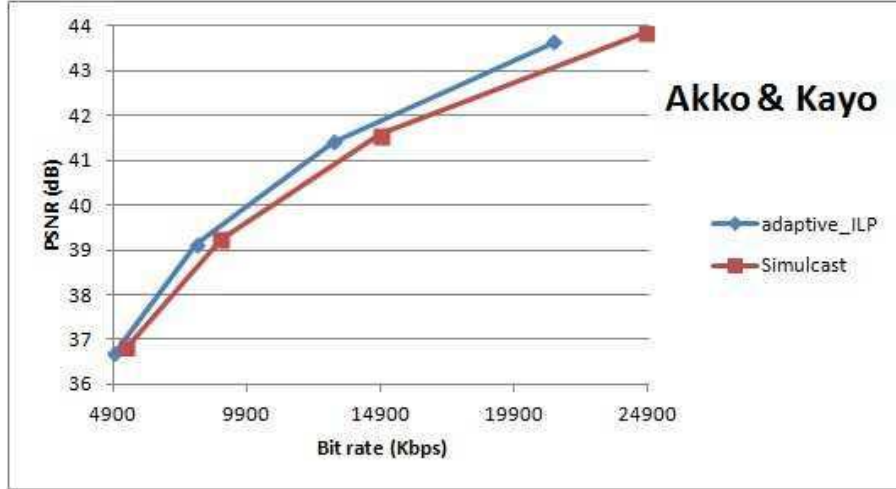
Sequence name	Bit rate percentages at lower frame rates (%)					Avg. extraction time per frame(ms)
	v0 = 1 v1 = 1/2	v0 = 1 v1 = 1/4	v0 = 1/2 v1 = 1/2	v0 = 1/2 v1 = 1/4	v0 = 1/4 v1 = 1/4	
Akko& kayo	90.91	80.5	82.53	72.12	62.33	0.19
Crowd	90.67	79.89	79.84	69.07	57.06	0.32
Flamenco	90.28	79.39	78.35	67.46	54.88	0.21
Objects	93.02	86.01	86.52	79.52	73.51	0.16
Balloons	90.22	80.93	79.51	70.22	60.36	0.38

### 7.3. CGS-temporal scalability

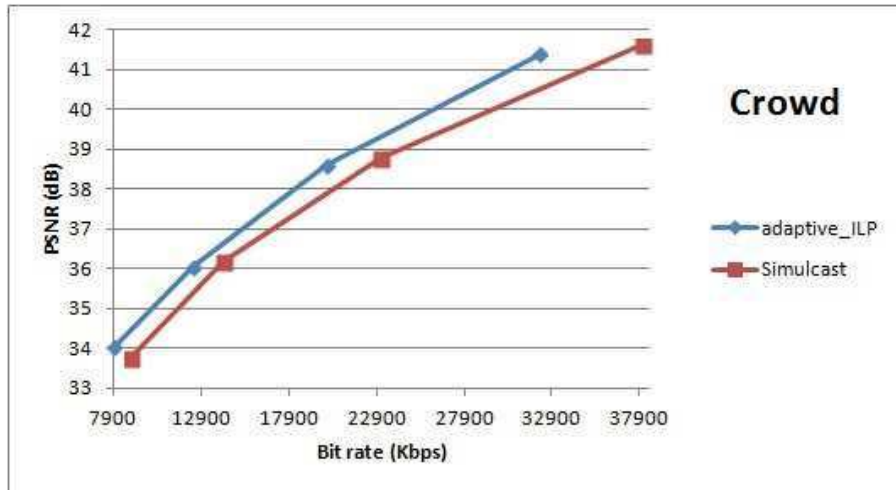
In this simulation setup, the R-D performance of the proposed SMVC method is compared with the simulcast method for the case of CGS-temporal scalability. The prediction structure employed in the previous sections has been used for this simulation as well and the enhancement layer video is encoded at a frame rate of ‘N’ and the base layer is encoded at a frame rate of ‘N/2’.The simulation results are shown in Table XVII. Also the rate-distortion curves are shown in Fig.7.8 – Fig.7.11 for various test sequences.

**Table XVII:** R-D performance results for CGS-temporal scalability

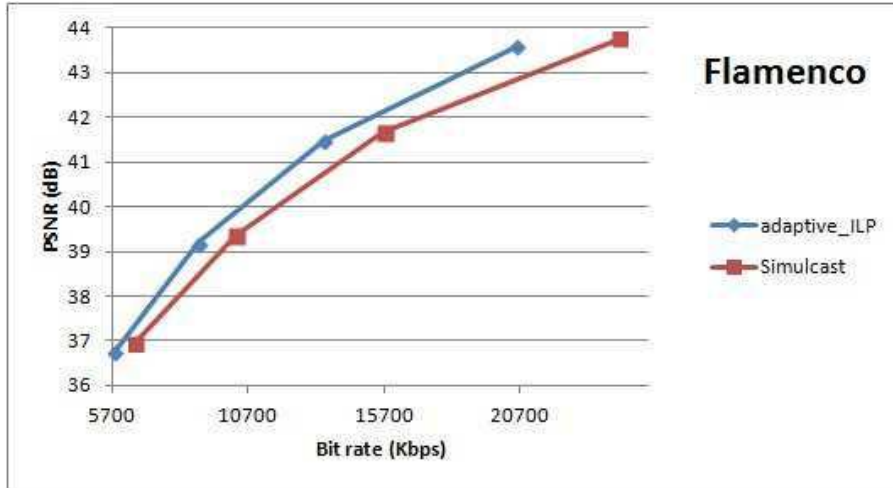
Sequence	QP	Proposed method		Simulcast method		Delta		Percentage bit rate Gain (%)
		Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)	
Crowd	BL:22, EL:20	32220.1	41.3885	38033.89	41.61	-5813.7	-0.2265	30.23
	BL:26, EL:24	20050.8	38.6165	23077.09	38.78	-3026.25	-0.1725	25.56
	BL:30, EL:28	12420.1	36.063	14139.26	36.196	-1719.11	-0.1335	23.46
	BL:34, EL:32	7958.7	34.0365	8846.56	33.755	-887.86	0.2815	20.14
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-0.06275</b>	<b>24.85</b>
Flamenco	BL:22, EL:20	20642.0	43.586	24383.71	43.756	-3741.64	-0.17	29.73
	BL:26, EL:24	13496.2	41.473	15695.85	41.649	-2199.57	-0.176	26.78
	BL:30, EL:28	8857.95	39.1955	10225.34	39.360	-1367.39	-0.165	25.13
	BL:34, EL:32	5786.54	36.7735	6500.03	36.93	-713.49	-0.1565	21.04
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-0.16687</b>	<b>25.67</b>
Akko & Kayo	BL:22, EL:20	21395.0	43.6355	24833.44	43.853	-3438.42	-0.2175	27.07
	BL:26, EL:24	13149.0	41.4115	14872.38	41.577	-1723.31	-0.166	22.21
	BL:30, EL:28	7996.16	39.114	8905.14	39.243	-908.98	-0.1295	19.23
	BL:34, EL:32	4930.78	36.7065	5364.22	36.843	-433.44	-0.137	15.56
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-0.1625</b>	<b>21.02</b>
Race	BL:22, EL:20	35824.7	41.8275	40941.04	42.002	-5116.32	-0.1745	25.43
	BL:26, EL:24	22997.4	39.2975	26149.12	39.451	-3151.7	-0.1535	24.43
	BL:30, EL:28	14700.3	36.827	16714.05	36.9795	-2013.74	-0.1525	24.17
	BL:34, EL:32	9499.96	34.4575	10705.35	34.612	-1205.39	-0.1545	23.66
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-0.15875</b>	<b>24.42</b>
<b>Overall average delta PSNR and percentage bit rate gain</b>							<b>-0.13771</b>	<b>23.99</b>



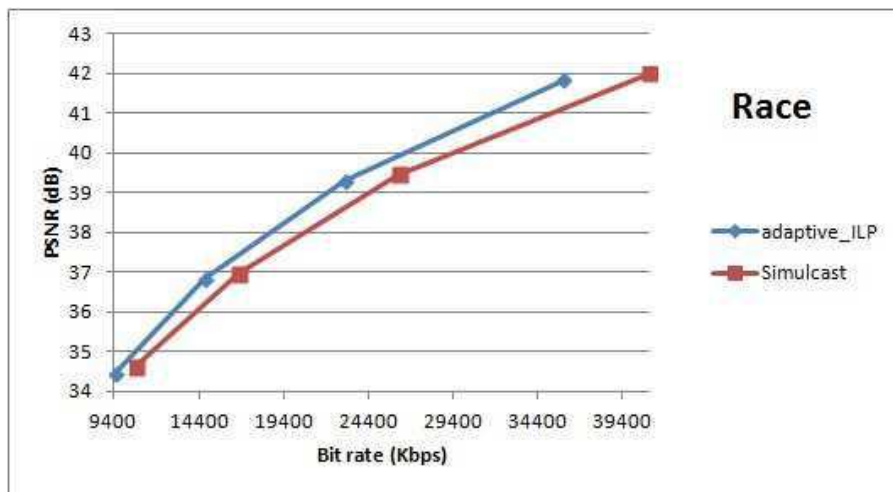
**Figure 7.8:** R-D curves for Akko & Kayo sequence: CGS-temporal scalability



**Figure 7.9:** R-D curves for Crowd sequence: CGS-temporal scalability



**Figure 7.10:** R-D curves for Flamenco sequence: CGS-temporal scalability



**Figure 7.11:** R-D curves for Race sequence: CGS-temporal scalability

The term adaptive\_ILP in the R-D curve graphs indicates the proposed scalable multi view video coding method and from the table, R-D curves, it could be noted that SMVC method provides around 25% reduction in bit rate with average PSNR reduction of only 0.1 dB.

The compression ratio and encoder speed timings (in terms of frames per second) are shown in Table XVIII. The formulae used to compute the bit rate of the original, compressed videos and compression ratio are same as that of spatial scalability scenario mentioned in Sec.6.2.2 and the reason mentioned in Sec.6.2.3 w.r.t increased original video bit rate holds good for this section as well. The performance numbers were measured on an Intel core-i5 laptop, with 4 GB RAM and clocked at 3.0 GHz. The multi-view video numbers (depicted in the last column of Table XVIII) are projected numbers for three views with estimated complexity increase of 50%. The justification behind the estimate mentioned in Sec.6.2.2 holds good for this section as well. Again, the frame rates depict the processing speed and is excluding the FILE Input / Output timings. Similar to the spatial scalability scenario, it could be noted that there is no multi-threading implementation and application is using single core. The frame rates could be increased by processing individual views in separate thread as all the tools of the proposed SMVC framework in the current view depends only on the corresponding view in the reference layer. Also the usage of “early termination” algorithms will help the encoder in arriving at the best prediction option without performing an exhaustive evaluation of all the inter-layer and intra-layer prediction options.

**Table XVIII:** Compression ratio and encoder processing speed for CGS-temporal scalability

Sequence name	QP	Bit rate (original video) (Bytes per sec)	Bit rate (compressed video) (Bytes per sec)	Compression ratio	Encoder frame rate (frames per second) for stereo video	Encoder frame rate (frames per second) for multi-view video
Crowd	BL:22, EL:20	55296000	4027512.5	14	13.39	6.70
	BL:26, EL:24	55296000	2506350	22	10.16	5.08
	BL:30, EL:28	55296000	1552512.5	36	10.38	5.19
	BL:34, EL:32	55296000	994837.5	56	10.59	5.30
Flamenco	BL:22, EL:20	55296000	2580250	21	14.28	7.14
	BL:26, EL:24	55296000	1687025	33	10.47	5.24
	BL:30, EL:28	55296000	1107243.75	50	10.61	5.30
	BL:34, EL:32	55296000	723317.5	76	10.80	5.4
Akko & Kayo	BL:22, EL:20	55296000	2674375	21	15.12	7.56
	BL:26, EL:24	55296000	1643625	34	10.185	5.09
	BL:30, EL:28	55296000	999520	55	10.43	5.21
	BL:34, EL:32	55296000	616347.5	90	10.70	5.35
Race	BL:22, EL:20	55296000	4478087.5	12	13.04	6.52
	BL:26, EL:24	55296000	2874675	19	9.33	4.67
	BL:30, EL:28	55296000	1837537.5	30	9.645	4.82
	BL:34, EL:32	55296000	1187495	47	9.945	4.97

## **Chapter 8**

### **Error robustness**

This chapter explains the error robustness provided by the proposed scalable multi view video coding specification as problems arise when multi-view video data shall be transmitted over error-prone channels. This is an important scenario, since 3D TV and Free view point TV (FTV) could be broadcasted over terrestrial or satellite channels, which may lead to errors. Furthermore, streaming multi-view data over the internet is also a realistic possibility, which is likely to involve packet loss. Thus, transmission of compressed video content over transmission medium such as wireless channels is severely affected by packet loss. In this chapter an over view of the existing error robustness methods applicable for H.264 MVC have been presented followed by the proposed method.

#### **8.1. Need for error concealment**

The bit-stream errors that can be classified into two types,

- random bit
- burst errors

Since H.264 MVC has a spatial-temporal block-based structure, the produced bit-streams are very sensitive to transmission errors leading to spatial and temporal error propagation. Spatial error propagation occurs when there is a loss of synchronization in predictive or entropy decoding. Temporal and inter-view error propagation result due to the use of motion compensation techniques between temporal and inter-view frames respectively. The H.264 MVC standard only considers the proper definition of the syntax and semantics of the bit-stream and it does not give any solution for erroneous bit-streams. It assumes that the lower network layer has the capability to detect and drop corrupted packets so that the decoder is only presented with intact packets. Therefore, error resilient coding is required to limit the propagation of visual artifacts while error concealment can be adopted to minimize the visual artifacts caused by the lost/corrupted slices.

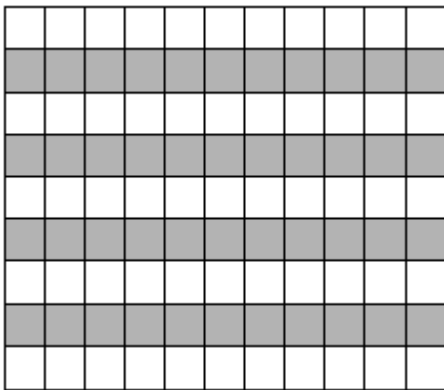
Error resilient mechanisms are introduced at the encoder to make the transmitted video bit-streams more robust to potential errors and to facilitate error concealment at the decoder. The error resilient schemes adopted by H.264 MVC to mitigate the effect of packet loss are:

- Slice coding that limits the spatial error propagation
- Insertion of regular Intra coded frames to limit the temporal propagation of the error

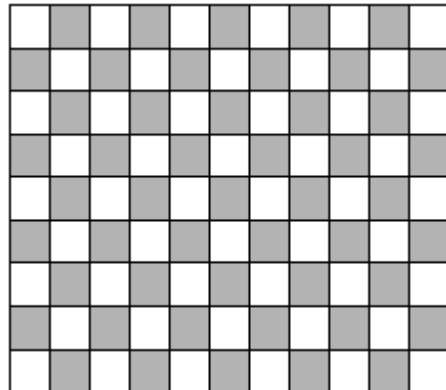


- Flexible Macro-block Ordering (FMO) that aids in error concealment

In [74], an overview of the performance of slice coding together with the effect of the cyclic-Intra coded period for H.264/MVC bit-streams has been presented. Slice coding is also a requirement for most concealment methods since it prevents an entire picture from being lost [75]. FMO allows the Macro-blocks (MBs) to be grouped in slice-groups and each of them can be partitioned into one or more slices. Various FMO types are available [76], but the most efficient is the dispersed type FMO, where consecutive MBs are transmitted in different slice-groups to protect the neighborhood. FMO interleaving mode and FMO checker-board mode are shown in Fig.8.1.a, Fig.8.1.b respectively, with two slices for a frame. The blocks with the same color will be grouped into one slice. FMO uniformly scatters possible errors to the whole frame to avoid error accumulation in a limited region [77]. Both slice coding and FMO increase the probability that a corrupted MB has distortion-free neighbors which can be used to aid concealment.



**Figure 8.1.a:** FMO interleaving mode



**Figure 8.1.b:** FMO Checker board mode

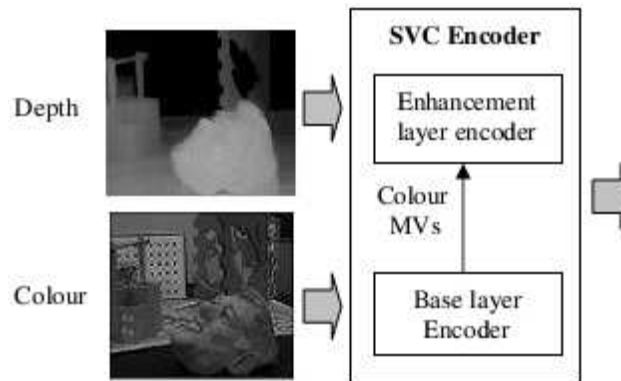
Error concealment methods are used at the decoder to estimate the lost information by taking advantage of the inherent correlation that exists between the missing MB and the neighborhood MBs. For multi-view concealment, each view can be concealed separately using only the single-view spatial and temporal error concealment techniques. However, for better error concealment, the neighborhood view-point frame together with the depth video can be exploited since this has also a high correlation with the corrupted frame.

## 8.2. Error concealment schemes in H.264 MVC

### 8.2.1. Depth image based error-concealment schemes

In [78], the authors propose a frame concealment method using shared motion vectors (MVs) between the color and depth data. A layered architecture similar to H.264 SVC (Scalable video coding) has been used, where the

colour information is coded in the base layer and the depth information is coded in the enhancement layer, as shown in Fig.8.2.



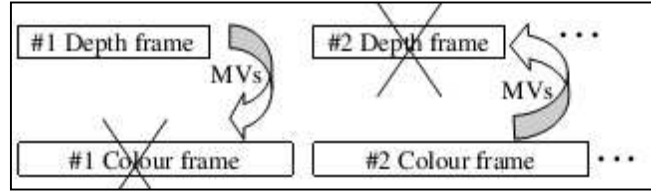
© 2008, IEEE

**Figure 8.2:** Depth image based concealment scheme [78]

For the color image sequence, one MV is generated for each macroblock (MB) using the motion estimation process. Two encoding modes are used to perform rate-distortion (R-D) optimization during depth map coding, namely ‘MB skip’ and ‘Motion compensation’. As large numbers of MBs in the depth image sequence have a uniform texture without high frequency components, the ‘MB skip’ mode is used to increase the compression efficiency. The ‘Motion compensation’ mode does not perform motion estimation. Instead, it reuses the MVs of the corresponding color image MBs to predict the current depth image at the enhancement layer. Then, the difference between the predicted and original frames is transform coded. The ‘Motion compensation’ mode is performed faster than conventional motion compensation modes as motion estimation does not take place during depth map coding. Finally, an R-D optimized encoding mode is selected to encode the MBs of the depth image. Hence, the global SVC bit-stream consists of both base and enhancement layer units, which include headers, shared MVs and coded residual texture data.

The overhead added due to the depth information needs to be kept at a smaller percentage of its corresponding color video. Due to the areas where the motion of color and depth map is not highly correlated, the proposed MV sharing method imposes a bit budget penalty during depth map coding. However, high quality depth maps can be obtained by adjusting the depth information percentage within an acceptable percentage of color bitrate. Thus, the average depth bitrate is kept below 25% of its corresponding color video bitrate.

If the color frame is received corrupted, the MVs from the corresponding uncorrupted depth frame are used to form the concealed frame, and vice versa as shown in Fig.8.3.



© 2008, IEEE

**Figure 8.3:** Error concealment scheme using depth and color image [78]

If a color video frame is missing due to packet losses, the MVs of the lost frame can be recovered using the correctly received corresponding depth video frame. Similarly, when a depth frame is lost, then the MVs can be recovered from the uncorrupted corresponding color video frame. The recovered MVs from the corresponding view are used to predict the current frame. If both corresponding frames are lost, then conventional single-view concealment algorithms are used to recover the lost frame. The authors in [79] consider the correlation between the color video and the depth video to propose a temporal error concealment technique for the lost MB. Another way to conceal a Multi View Video (MVV) is to use view interpolation [80] to obtain the lost frames. However, since MVV can be encoded efficiently without using the depth information, these techniques are not feasible as they require more processing resources and bandwidth.

### 8.2.2. Disparity vector based frame-loss concealment schemes

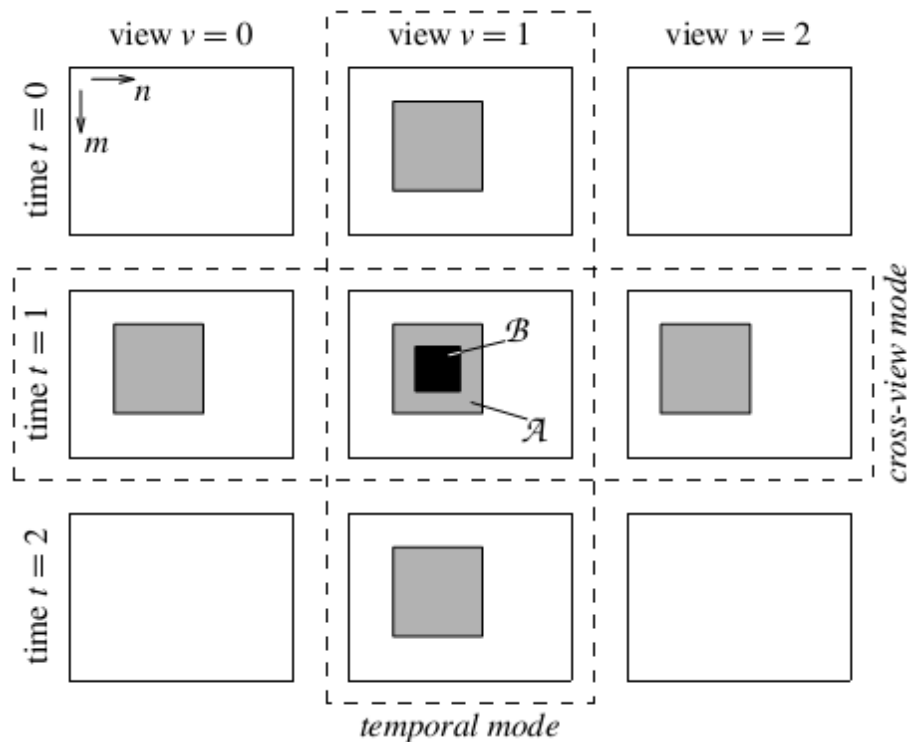
In [81], the authors calculate a global Disparity Vector (DV), for the inter-view referenced frame, relative to the base view and this is transmitted with the anchor frames. When a frame is lost, the corresponding MBs in the dependent frame are located using the global DV. For each MB, the mode and the MVs are copied and Motion compensation (MC) is used to generate the concealed frame. If the matching MB does not contain MVs, then depending on the frames' type, it is either spatially concealed or a MV is estimated for it.

In [82], the authors propose another method for full frame concealment where, following a number of steps, different algorithms are used to estimate the MVs and DVs for each MB from the DV or the MV or both fields of the source frame. Then, a median filter is applied to filter both the MV and DV fields, to fill the empty spaces and to filter irregularities. These vectors are used with motion and disparity compensation to form the concealed frame. Finally, the resulting picture is filtered again using a median filter to fill the empty regions.

### 8.2.3. Disparity vector based slice-loss concealment schemes

The methods presented in [81-82] consider the loss of a whole frame but the effect of transmission errors can be reduced by transmitting the frames in smaller slices. This is obtained through slice coding that result in slices

with a smaller probability of error [75]. Since only the corrupted parts of the image will be concealed, the reconstructed quality of the frame improves. This increases the bandwidth required for MVVs [75] but provides better concealment since higher fidelity neighborhood data is available to predict lost regions. In literature, algorithms that consider concealment on multi-view slices can be found in [83-86]. The authors in [83] consider a four-dimensional frequency selective extrapolation process which exploits the surrounding information within the same image and information from temporal and view-point neighboring frames to estimate the missing samples in the corrupted area. The four-dimensional extrapolation scheme is shown in Fig.8.4. Basically, a multi-view video signal can be described as a four-dimensional data volume. Two dimensions 'm' and 'n' are necessary for the spatial directions (vertical and horizontal) within a single image. The third dimension 't' denotes the time axis, and the fourth dimension 'v' the camera view number. The loss in the slice information could be concealed using any of the same image / same time / same access unit information.



© 2008, IEEE

**Figure 8.4:** 4-D arrangement of MVV data [83]

In [84] and [85] the authors consider a method that identifies the corresponding region in the reference frame through feature points. This region is used together with the boundary pixels in a weighted sum to obtain the replacing MB. However, the latter two methods are highly dependent on the quality of the reference frames. They

only consider the pixel values within the reference frames and they do not make use of MVs or DVs which are useful for better concealment especially in highly dynamic scenes.

#### **8.2.4. Concealment using decoder side ME**

In [86], the authors estimate the lost MBs by estimating the MVs and DVs from the neighborhood temporal and inter-view frames, respectively. The outer boundary of the lost MB is considered and a full search for the replacing MB that minimizes the boundary distortion error is searched in the temporal and the view-point frames using the Decoder Motion Vector Estimation (DMVE) technique. This method is adopted from single-view concealment [87]. All the MVs and DVs which give a small distortion error are used for motion compensation and the formed MBs are weighted to form the replacing MB. This technique provides good concealment however the complexity of the decoder is drastically increased since the decoder must search for the optimal vectors for the replacing MBs.

#### **8.2.5. Forward error correction based concealment schemes**

Other work, such as [88-89], uses Forward Error Correction (FEC) schemes to introduce redundancies in the code words to ease their correction. These are used to form unequal error protection on different MVV elements such that better protection is given to more important data. In [88], MBs are classified into slice-groups by examining their relative significance to the video and more important MBs are transmitted with better protection, by using the explicit type FMO [76]. In [89] the unequal error protection is formed by protecting different frame types with different levels of protection. Intra coded frames are the most protected, followed by temporal predicted frames, followed by inter-view predicted frame. Although this provides good error resilience, it increases the transmission bandwidth. If this is compensated for by decreasing the video bandwidth to accommodate it, the quality of the uncorrupted frames is reduced. With these schemes simple error concealment techniques can be used.

### **8.3. Proposed error-concealment scheme using SMVC**

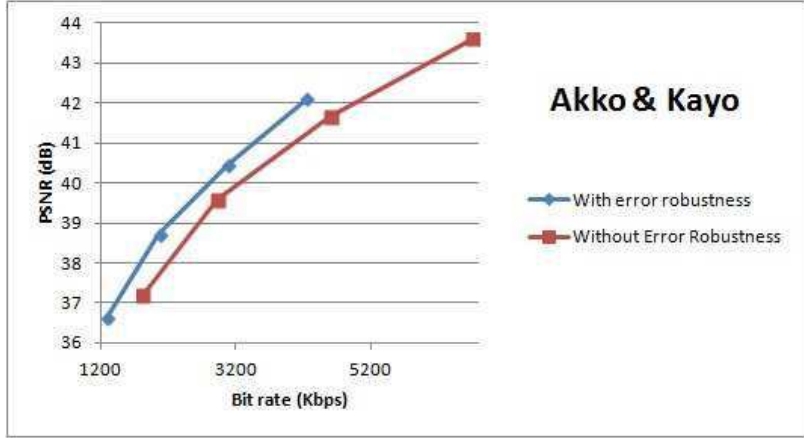
The methods presented in Sec.8.2.1, Sec.8.2.4, Sec.8.2.5 are computationally expensive and the proposed error concealment could be considered as the category of disparity vector based methods as depicted in Sec.8.2.2, Sec.8.2.3. The proposed error concealment method works as follows; in the event of packet loss, the decoder will use the data present in the reference layers to conceal the data.

### 8.3.1. Error concealment for CGS case

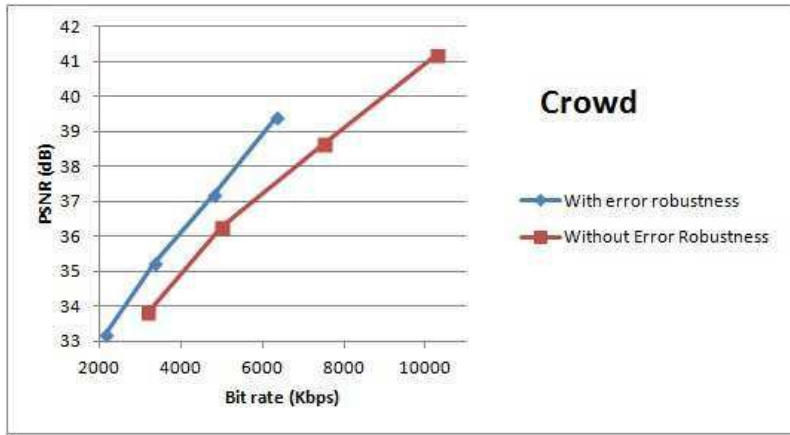
In this simulation setup, the error robustness performance of the proposed SMVC method is evaluated for CGS scalability scenarios. The bitstream extractor tool provided by the JMVC reference software [55] has been modified to drop the NAL packets of the highest temporal layer frames of the enhancement layer. The decoder conceals the error by copying the motion vectors, reference indices, MB type information and residual data from the reference layer and performs motion compensation. The R-D performance are shown in Table XIX and in Fig.8.5 – Fig.8.8, it could be seen that approximately 35% savings in bit rate could be achieved by this selective dropping of packets with an average PSNR reduction of around 1 dB. The reconstructed frames for with and without error scenarios are shown in Fig.8.9 – Fig.8.12 and it could be noted that subjective quality wise, the reconstructed frames in the erroneous scenario are on par with the without error scenario case.

**Table XIX:** R-D performance results error robustness of CGS scalability

Sequence	QP	With Errors in the bit stream		Original bit stream		Delta		Percentage bit rate gain (%)
		Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)	
Crowd	BL:22, EL:20	6342.39	39.40	10294.1	41.203	-3951.78	-1.80	38.39
	BL:26, EL:24	4784.48	37.20	7492.13	38.664	-2707.65	-1.45	36.14
	BL:30, EL:28	3327.5	35.26	4980.07	36.281	-1652.57	-1.02	33.18
	BL:34, EL:32	2140.5	33.17	3178.51	33.855	-1037.92	-0.66	32.65
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-1.234</b>	<b>35.09</b>
Flamenco	BL:22, EL:20	3697.81	41.57	5995.83	43.337	-2298.02	-1.763	38.33
	BL:26, EL:24	2730.62	39.63	4424.49	41.307	-1693.87	-1.677	38.28
	BL:30, EL:28	2019.22	37.79	3210.7	39.179	-1191.48	-1.381	37.11
	BL:34, EL:32	1402.52	35.71	2217.5	36.756	-814.98	-1.04	36.75
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-1.465</b>	<b>37.62</b>
Akko & Kayo	BL:22, EL:20	4244.61	42.09	6721.57	43.612	-2476.96	-1.513	36.85
	BL:26, EL:24	3077.94	40.46	4606.51	41.654	-1528.57	-1.188	33.18
	BL:30, EL:28	2049.74	38.73	2927.26	39.579	-877.52	-0.847	29.98
	BL:34, EL:32	1285.21	36.63	1820.62	37.19	-535.41	-0.557	29.41
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-1.02625</b>	<b>32.35</b>
Race	BL:22, EL:20	5926.8	40.69	9129.67	41.929	-3202.87	-1.239	35.08
	BL:26, EL:24	3964.94	38.57	6056.85	39.639	-2091.91	-1.062	34.54
	BL:30, EL:28	2691.67	36.50	4002.43	37.38	-1310.76	-0.877	32.75
	BL:34, EL:32	1701.77	34.36	2505.69	35.04	-803.92	-0.671	32.08
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-0.96225</b>	<b>33.61</b>

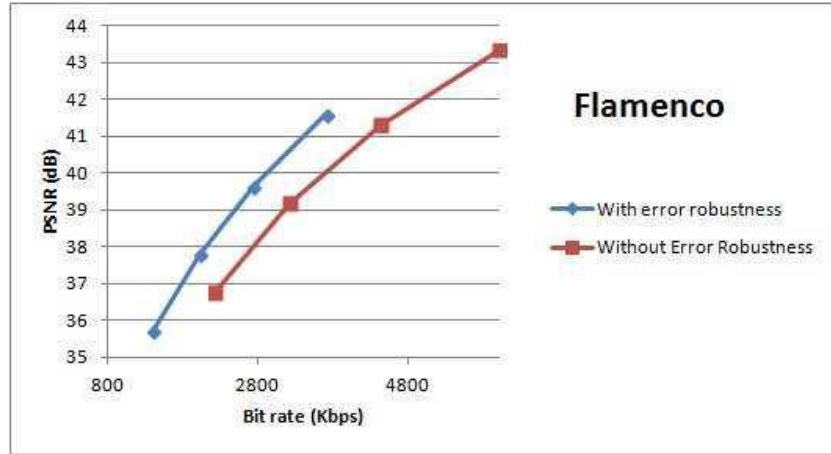


**Figure 8.5:** R-D curves for Akko & Kayo sequence: CGS error robustness

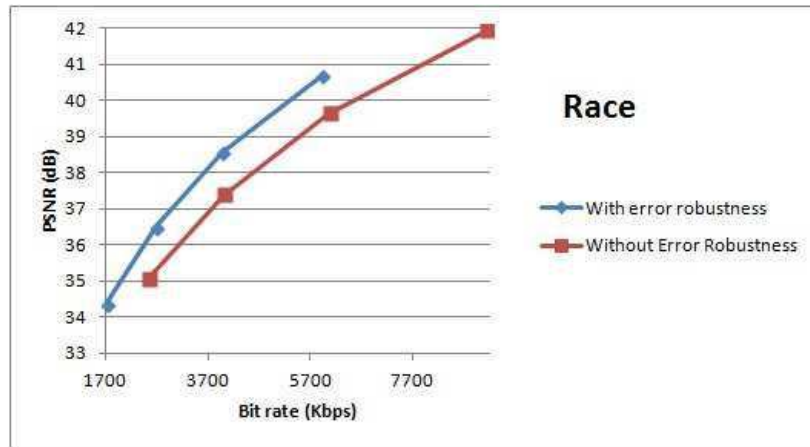


**Figure 8.6:** R-D curves for Crowd sequence: CGS error robustness





**Figure 8.7:** R-D curves for Flamenco sequence: CGS error robustness



**Figure 8.8:** R-D curves for Race sequence: CGS error robustness



**Figure 8.9:** CGS Subjective quality – Akko & Kayo: Originally encoded (PSNR – 42.9 dB) on the left and concealed frame on the right (PSNR – 38.6 dB)



**Figure 8.10:** CGS Subjective quality – Crowd: Originally encoded (PSNR – 40.7 dB) on the left and concealed frame on the right (PSNR – 35.5 dB)



**Figure 8.11:** CGS Subjective quality – Flamenco: Originally encoded (PSNR – 42.8 dB) on the left and concealed frame on the right (PSNR – 37.8 dB)



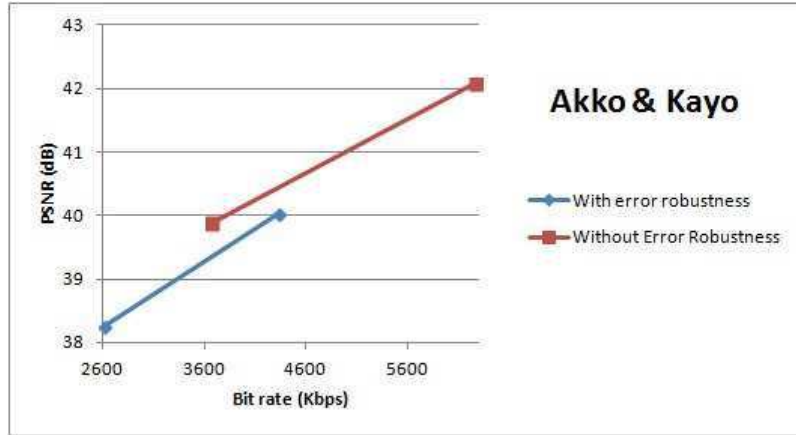
**Figure 8.12:** CGS Subjective quality – Race: Originally encoded (PSNR – 41.58 dB) on the left and concealed frame on the right (PSNR -39.23 dB)

### 8.3.2. Error concealment for SS case

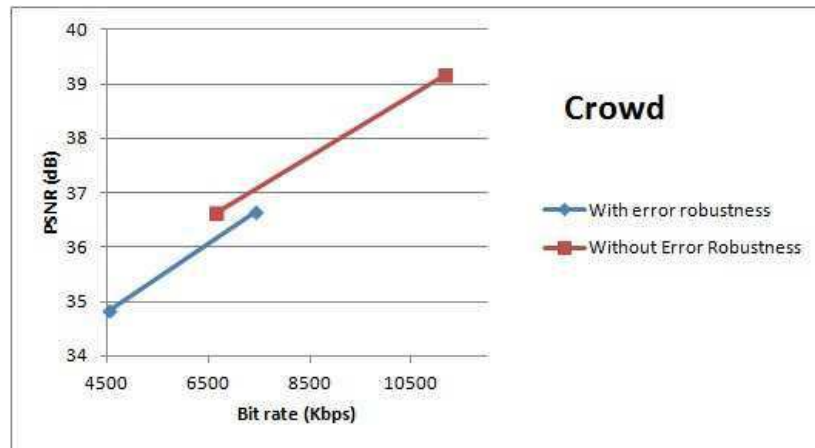
This simulation setup is same as that of mentioned in Sec.8.3.1 with the exception that the enhancement layer frames are of higher resolution when compared to the reference layer frames. The R-D performance are depicted in Table XX as well as in Fig.8.13 – Fig.8.16, it could be seen that approximately 35% savings in bit rate could be achieved by this selective dropping of packets with an average PSNR reduction of around two dB. The reconstructed frames for with and without error scenarios are shown in Fig.8.17 – Fig.8.20 and it could be noted that subjective quality wise, the reconstructed frames in the erroneous scenario are on par with the without error scenario case.

**Table XX : R-D performance results error robustness of spatial scalability**

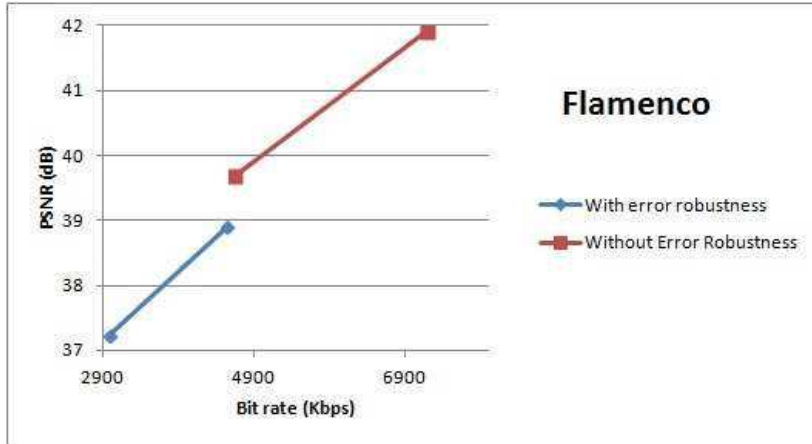
Sequence	QP	With Errors in the bit stream		Original bit stream		Delta		Percentage bit rate gain (%)
		Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)	Bit rate (Kbps)	PSNR (dB)	
Crowd	BL:22, EL:20	12081.15	38.391	18656.28	41.928	-6575.13	-3.537	35.24
	BL:26, EL:24	7429.14	36.649	11164.16	39.171	-3735.02	-2.522	33.46
	BL:30, EL:28	4550.53	34.853	6651.78	36.635	-2101.25	-1.782	31.59
	BL:34, EL:32	2725.34	32.843	3908.09	34.125	-1182.75	-1.282	30.26
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-2.281</b>	<b>32.64</b>
Flamenco	BL:22, EL:20	7006.06	40.44	11260.33	43.953	-4254.27	-3.513	37.78
	BL:26, EL:24	4526.67	38.892	7193.3	41.905	-2666.63	-3.013	37.07
	BL:30, EL:28	2978.23	37.227	4647.84	39.694	-1669.61	-2.467	35.92
	BL:34, EL:32	1887.54	35.281	2879.23	37.206	-991.69	-1.925	34.44
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-2.722</b>	<b>36.30</b>
Akko & Kayo	BL:22, EL:20	6963.46	41.673	10501.08	44.194	-3537.62	-2.521	33.69
	BL:26, EL:24	4328.37	40.024	6274.5	42.076	-1946.13	-2.052	31.02
	BL:30, EL:28	2615.76	38.256	3675.5	39.879	-1059.74	-1.623	28.83
	BL:34, EL:32	1507.7	36.119	2104.62	37.454	-596.92	-1.335	28.36
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-1.8827</b>	<b>30.47</b>
Race	BL:22, EL:20	8955.66	39.688	13941.98	42.473	-4986.32	-2.785	35.76
	BL:26, EL:24	5401.74	38.044	8243.24	40.12	-2841.5	-2.076	34.47
	BL:30, EL:28	3373.22	36.246	5012.71	37.811	-1639.49	-1.565	32.71
	BL:34, EL:32	2049.08	34.259	2981.8	35.459	-932.72	-1.2	31.28
<b>Average delta PSNR and percentage bit rate gain</b>							<b>-1.9065</b>	<b>33.56</b>



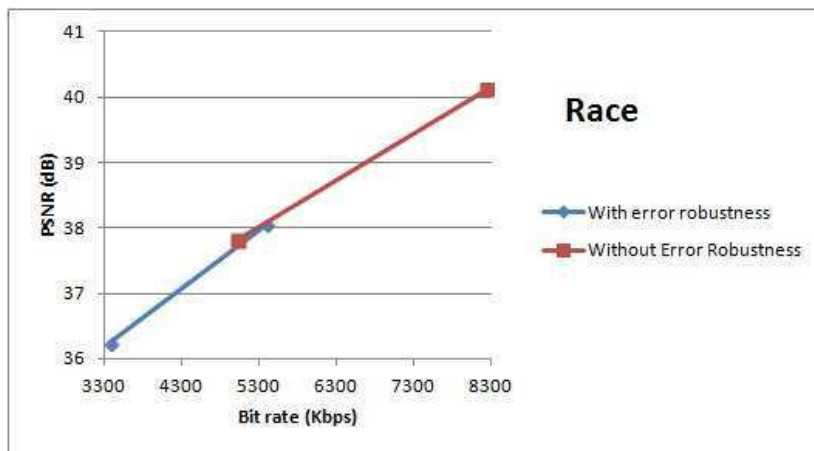
**Figure 8.13:** R-D curves for Akko & Kayo sequence: SS error robustness



**Figure 8.14:** R-D curves for crowd sequence: SS error robustness



**Figure 8.15:** R-D curves for Flamenco sequence: SS error robustness



**Figure 8.16:** R-D curves for Race sequence: SS error robustness



**Figure 8.17:** SS Subjective quality – Akko & Kayo: Originally encoded frame (PSNR – 43.38 dB) on the left and concealed frame (PSNR – 37.6 dB) on the right



**Figure 8.18:** SS Subjective quality – Flamenco: Originally encoded frame (PSNR –43.5 dB) on the left and concealed frame (PSNR – 35.7 dB) on the right



**Figure 8.19:** SS Subjective quality – Crowd: Originally encoded frame (PSNR – 40.9 dB) on the left and concealed frame (PSNR – 32.25 dB) on the right



**Figure 8.20:** SS Subjective quality – Race: Originally encoded frame (PSNR – 42.4 dB) on the left and concealed frame (PSNR – 37.5 dB) on the right

## **Chapter 9**

### **Conclusions and future work**

The aim of the thesis was to introduce a fully scalable and backward-compatible multi view video coding framework based on H.264 multi view video coding specification. There are a number of novel coding tools such as inter-layer intra prediction, inter-layer motion prediction, inter-layer residual prediction and disparity compensated prediction have been introduced with emphasis on single loop decoding and parallel decoding of layers. The thesis was intended for multi view video use cases such as 3DTV and immersive 3D video conferencing and more specifically it enables interoperability between various multi view video devices connected with each other over a number of heterogeneous networks. Many of the standardization bodies such as DVB have planned for 3DTV deployments in two phases; the first phase is termed as ‘frame-compatible’ phase where the existing 2D infrastructure including the 2D compression and decompression schemes will be reused. Second phase is termed as ‘service compatible’ where the state-of-the-art multi view video processing infrastructure and multi view video compression schemes will be used for the introduction of more advanced features of 3D. The contributions of the thesis are summarized below.

For the first phase, the thesis proposed a novel method using which the video frames from left and right cameras could be packed in to two separate tiles. The resultant frame will be encoded using HEVC compression scheme and additionally in-loop filters will be turned off across tile boundary. The intuition behind this is that due to independent nature of the left and right view data, the loop filters will only introduce noise instead of smoothening the pixels present across the tile border. On the other hand, if the in-loop filters are turned off at the tile boundary, complete encoding/decoding of frames could happen in parallel without affecting the rate-distortion performance. Experimental results show that the proposed scheme enables parallel encoding/decoding of frames and improves the processing speed by approximately twice while maintaining the same rate-distortion performance. This will be extremely useful in ‘latency dominant’ scenarios such as low delay stereo video encoding/decoding.

A comprehensive study of various options including spatial interleaving, temporal interleaving, wavelet coding based and hybrid coding based methods have been carried out. The purpose of the study was to analyze these methods from a number of different perspectives such as compression efficiency, backward compatibility, stereoscopic 3D support, re-use of existing infrastructure, single loop decoding, support for various scalability options and choose the best among the above-mentioned methods for the realization of scalable multi view video coding compression scheme. In order to identify the best compression scheme to start with, simulations have been setup to benchmark the rate-distortion performance of the various above-mentioned methods and results show that



H.264 multi view video coding (which belong to the hybrid coding based scheme) performs better when compared to all the other methods for the stereoscopic video content and also it meets a number of scalable multi view video coding requirements compared to others. This clearly indicate that different multi view capable devices cannot be serviced by the first phase implementation (as only spatial interleaving scheme is supported in the first phase) and there is a need for more advanced multi view video compression. H.264 multi view video coding will form the starting point for the realization of this advanced multi view video compression scheme.

For the second phase, there are a number of contributions in this thesis; the existing temporal scalability mechanism in the H.264 multi view video coding has been studied and two improvements have been proposed. First improvement is that a new state-of-the-art prediction structure, which can generate a fully scalable bitstream has been proposed. The deficiency of the existing prediction structure which is based on "disposable view components" is that temporal scalability requirements could not be met to the fullest and it is shown that the proposed prediction structure will overcome this limitation. Simulation results show that a number of subsequences could be extracted by using the proposed prediction structure with minimal computational overhead at the media aware network server. The application of the proposed prediction structure is that when there is a drop in available transmission bandwidth, the multi view devices could still be serviced and connected to the network by discarding the higher layer temporal frames.

The second suggested improvement is to the use temporal identifier information in the reference picture list construction process for better rate-distortion performance compared to existing H.264 MVC specification. In the current specification, temporal identifier information is not used in the reference picture list construction process and instead used only in the sub bitstream extraction process. Simulation results show that for full HD resolution (1920x1080) video coded at 30 frames per second with three temporal layer scheme, around 300 Mbps savings in bit rate could be achieved by using the proposed method. Only predictive slices with single reference picture list has been used for the simulation and when bi-predictive slices are used, there will be more savings in bit rate. Hence the simulation results show the lower bound of achievable savings in bit rate. This will be extremely useful in scenarios where the number of views will be extremely high as in the case of free view point TV. The proposed method also ensures that the state of the decoded picture buffer at the end of 'default' reference picture list construction process will be same in encoder and decoder even after frame loss, which will help the decoder in coping with the transmission errors.

The spatial scalability and coarse grain bit rate scalability (CGS) options were also introduced by this thesis in addition to the existing view scalability and temporal scalability. Simulations have been setup to benchmark the rate-distortion performance of the introduced tools with simulcast option for VGA resolution (640x480) videos coded at 30 frames per second and results show that the proposed method provides around 20% reduction in bit

rate with a max PSNR reduction of 0.23 dB in the case of coarse grain bit rate scalability and provides around 10% reduction in bit rate with a max PSNR reduction of 0.07 dB in the case of spatial scalability. Again only predictive slices are used for the simulation and results show the lower bound of achievable gain in bit rate. The method proposed in the thesis also provides the flexibility that spatial and CGS layers could be coded in separate layers of a given bitstream. This will ensure that a number of multi view video capable devices with different spatial resolutions could be connected with each other over heterogeneous networks, just by using a single encoded bitstream.

The thesis also showed that combined scalability options such as spatio-temporal and CGS-temporal could be realized to enhance the number of supported devices. The proposed method in the thesis provides superior error robustness, as it uses the coded information from the reference layers to conceal the errors in the target layer. The error robustness could be used to save bits by selective dropping of packets and will be useful when there is reduction in available transmission bandwidth. Simulation results corresponding to VGA resolution (640x480) videos coded at 30 frames per second with three temporal layer scheme, show that using this method around 35% reduction in bit rate could be achieved with a max PSNR reduction of around 2dB. Subjective quality results also show that the error in the reconstructed frames is not perceivable and the proposed method could be especially useful in low bit rate, low delay encoding scenarios such as 3D video conferencing.

The potential future work could be summarized as follows; even though the proposed framework supports parallel encoding/decoding of layers, the corresponding software changes have not been made. Also, since the JM reference software supports stereoscopic video encoding/decoding only, all the simulations have been limited to stereoscopic video content. The thesis is based on H.264 multi view video coding which is the current state-of-the-art in the industry, but recently a high efficiency video coding based multi view coding (MV-HEVC) has been standardized. The feasibility of extending the proposed method in the thesis for MV-HEVC could be studied in future along with the corresponding software changes.

## Bibliography

- [1] Anil Fernando., Stewart T. Worrall., and Erhan Ekmekciođlu. 3DTV: Processing and Transmission of 3D Video Signals. John Wiley & Sons, 2013.
- [2] Masayuki Tanimoto., Mehrdad Panahpour Tehrani., Toshiaki Fujii., and Tomohiro Yendo., "Free-Viewpoint TV", IEEE Signal Processing Magazine, vol.28, no.1, pp.67-76, January 2011.
- [3] Masayuki Tanimoto., Mehrdad Panahpour Tehrani., Toshiaki Fujii., and Tomohiro Yendo., "FTV for 3-D Spatial Communication", Proceedings of the IEEE, Vol. 100, No. 4, pp. 905-917, April 2012.
- [4] Mehrdad Panahpour Tehrani., Takanori Senoh., Makoto Okui., Kenji Yamamoto., Naomi Inoue., and Toshiaki Fujii., "Introduction of Super Multiview Video Systems for Requirement Discussion", ISO/IEC JTC1/SC29/WG11 M31052, Geneva, CH, Oct. 2013.
- [5] LG 3DTV with Disney 3D movies available on rental.<http://www.engadget.com/2012/10/06/lg-smart-tv-platform-disney-3d-movie-rentals/>.
- [6] Thomas Schierl., Thomas Stockhammer., and Thomas Wiegand., "Mobile video transmission using scalable video coding." IEEE Transactions on Circuits and Systems for Video Technology, Vol.17, No. 9, pp. 1204-1217, 2007.
- [7] Eisert, Peter., "Immersive 3D video conferencing: challenges, concepts, and implementation." Visual Communications and Image Processing, pp.69-79, 2003.
- [8] Fujii, Toshiaki., Kensaku Mori., Kazuya Takeda., Kenji Mase., Masayuki Tanimoto., and Yasuhito Suenaga., "Multipoint measuring system for video and sound-100-camera and microphone system." In IEEE International Conference on Multimedia and Expo (ICME), pp. 437-440, 2006.
- [9] Jens-Rainer Ohm., "Stereo/multiview video encoding using the MPEG family of standards." In Electronic Imaging'99, International Society for Optics and Photonics, pp. 242-253, 1999.
- [10] Chen Ying., Ye-Kui Wang., Miska M. Hannuksela., and Moncef Gabbouj., "Single-loop decoding for multiview video coding." In IEEE International Conference on Multimedia and expo (ICME), pp. 605-608. June 2008.
- [11] A.Vetro., T.Wiegand., and G.J.Sullivan., "Overview of the Stereo and Multiview Video Coding Extensions of the H.264/MPEG-4 AVC Standard", Proceedings of the IEEE, vol.99, no.4, pp.626-642, 2011.

- [12] Gary J. Sullivan, “On the frame packing arrangement SEI message”, Study Group 16, VCEG Document. VCEG-AO12, 2010.
- [13] Digital Video Broadcasting (DVB), Frame Compatible Plano-stereoscopic 3DTV (DVB-3DTV), DVB Document, [http://www.dvb.org/technology/standards/a154\\_DVB-3DTV\\_Spec.pdf](http://www.dvb.org/technology/standards/a154_DVB-3DTV_Spec.pdf)
- [14] Andy Qusted, “Gearing up to deliver Wimbledon 3D”,  
[http://www.bbc.co.uk/blogs/bbcinternet/2011/06/gearing\\_up\\_to\\_deliver\\_wimbledo.html](http://www.bbc.co.uk/blogs/bbcinternet/2011/06/gearing_up_to_deliver_wimbledo.html) (Last accessed on 03/03/2016)
- [15] “Preparing 3D video with S3D metadata on HTC Evo”,  
<http://www.htcdev.com/devcenter/s3d> (Last accessed on 03/03/2016)
- [16] Ohji Nakagami and Teruhiko Suzuki, “On stereo 3D coding using frame packing arrangement SEI”, JCT-VC Document, JCTVC-I0058, 2012.
- [17] Ye- Kui Wang, AHG9: “Indication of frame-packed or interlaced video”, JCT-VC Document. JCTVC-K0119, 2012.
- [18] Xiaofeng Yang., Peiyu Yue., and Yuanyuan Zhang., “2D compatible frame packing stereo 3D video”, JCT-VC Document. JCTVC-K0116, 2012.
- [19] Yang W., Lu Y., Wu F., Cai J., Ngan KN., and Li S., “4-D Wavelet-Based Multiview Video Coding”, IEEE Transactions on Circuits and Systems for Video Technology. Vol.16, pp.1385–1396, 2006.
- [20] Garbas J.U., Fecker U., and Kaup A., “Wavelet-Based Multi-View Video Coding with Full Scalability and Illumination Compensation”. Proceedings of the 15th ACM International Conference on Multimedia, Germany, pp.751-754, September 2007.
- [21] Garbas J.U., and Kaup A., “Wavelet-Based Multi-View Video Coding with Spatial Scalability”. IEEE International Workshop on Multimedia Signal Processing (MMSP), Greece, pp.422-425, October 2007.
- [22] Garbas J.U., and Kaup A., “Enhancing Coding Efficiency in Spatial Scalable Multiview Video Coding with Wavelets”, IEEE International Workshop on Multimedia Signal Processing (MMSP), Brazil, pp.1-6, October 2009.

- [23] Xiong R., Xu J., Wu F., Li S., and Qin Zhang Y., “Spatial Scalability in 3D Wavelet Coding with Spatial Domain MCTF Encoder”. Proceedings of the Picture Coding Symposium (PCS), USA, pp.583-588, December 2004.
- [24] Xiong R., Xu J., Wu F., and Li S., “Studies on Spatial Scalable Frameworks for Motion Aligned 3D Wavelet Video Coding”. Proceedings of SPIE on visual communication and Image processing, Beijing, China, pp.189-200, July 2005.
- [25] Wiegand T., Sullivan G.J., Ohm J.R., and Luthra A. K., Introduction to the Special Issue on Scalable Video Coding—Standardization and Beyond. IEEE Transactions on Circuits and Systems for Video Technology, Vol.17, pp.1099-1102, 2007.
- [26] ISO/IEC JTC 1, “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), November 1994.
- [27] ITU-T, “Video coding for low bit rate communication, ITU-T Recommendation H.263, Version 1”: November 1995, Version 2: January 1998, Version 3: November 2000.
- [28] ISO/IEC JTC 1, “Coding of audio-visual objects - Part 2: Visual, ISO/IEC 14496-2 (MPEG-4 Visual)”, Version 1: April 1999, Version 2: February 2000, Version 3: May 2004.
- [29] Wei Yao, Zheng Guo Li., and Susanto Rahardja., “scalable video coding in a nutshell”, Synthesis journal, Vol.1, pp. 89-108, 2008.
- [30] Schwarz H., Hinz T., Marpe D., and Wiegand T., “Overview of the scalable H.264/MPEG4-AVC Extension”, Proceedings of International Conference in Image Processing (ICIP), USA, pp.161-164, October 2006.
- [31] ISO/IEC Joint Technical Committee, “Advanced video coding for generic audio-visual services, ITU-T Recommendation H.264 Amendment 3, ISO/IEC 14496-10/2005: Amd 3 - Scalable extension of H.264 (SVC)”, July 2007.
- [32] Ohm J.R., “Stereo/Multiview Encoding Using the MPEG Family of Standards”, Journal of Electronic Imaging, vol.1, pp.27-39, 1999.
- [33] ISO/IEC JTC1/SC29/WG11, “Requirements on Multi-view Video Coding v.5”, Doc.N7539, Nice, France, October 2005.

- [34] ISO/IEC JTC1/SC29/WG11, “Call for Proposals on Multi-view Video Coding”, Doc.N7327, Poznan, Poland, July 2005.
- [35] ISO/IEC Joint Technical Committee, “Advanced video coding for generic audio-visual services”, ITU-T Recommendation H.264 Amendment 3, ISO/IEC 14496-10/2005: “Annex H – Multi view extension of H.264 (MVC)”, March 2009.
- [36] Chen Y., Wang Y.K., Ugur K., Hannuksela M.M., Lainema J., and Gabbouj M., “The emerging MVC standard for 3D video services”. *EURASIP Journal on Advances in Signal Processing*, vol.1, pp.1- 13, 2009.
- [37] Anthony Vetro., Sehoon Yea., Matthias Zwicker., Wojciech Matusik., and Hanspeter Pfister., “Overview of Multiview Video Coding and Anti-Aliasing for 3D Displays”, *IEEE International Conference on Image Processing, USA*, pp. 17-20, October 2007.
- [38] ITU-T Recommendation H.264 (05/2003), “Advanced video coding for generic audio-visual services.”
- [39] ITU-T Recommendation H.264 (03/2005), “Advanced video coding for generic audio-visual services.”
- [40] ITU-T Recommendation H.264 (01/2012), “Advanced video coding for generic audio-visual services.”
- [41] ITU-T Recommendation H.264.2 (03/2005), “Reference software for H.264/AVC advanced video coding.”
- [42] ISO/IEC International Standard 14496-5:2001/Amd.6:2005, “Reference software for AVC and audio HE-AAC.”
- [43] K. Sühling (coordinator), “Joint model H.264/AVC reference software,” <http://iphome.hhi.de/suehring/tml/>.
- [44] G. J. Sullivan, P. N. Topiwala, and A. Luthra, “The H.264/AVC advanced video coding standard: overview and introduction to the fidelity range extensions,” *In Optical Science and Technology, the SPIE 49th Annual Meeting*, vol. 5558, pp. 454–474, 2004.
- [45] J. W. Woods, *Multidimensional Signal, Image and Video Processing and Coding*. Academic Press, 2006.
- [46] T. Wiegand, X. Zhang, and B. Girod, “Long-term memory motion-compensated prediction,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 70-84, February 1999.
- [47] A. Puri, X. Chen, and A. Luthra, “Video coding using the H.264/MPEG-4 AVC compression standard,” *Signal Processing: Image Communication*, vol. 19, no. 9, pp. 793-849, October 2004.
- [48] Tian, D., Hannuksela, M. M., and Gabbouj, M., “Sub-sequence video coding for improved temporal scalability,” *In IEEE International Symposium on Circuits and Systems (ISCAS)*, Vol 6, pp. 6074–6077, 2005.

- [49] Schwarz, H., Marpe, D., and Wiegand, T., "Analysis of hierarchical B pictures and MCTF", IEEE International Conference on Multimedia and Expo (ICME), Vol 1, pp. 1929–1932, 2006.
- [50] Hong, D., Horowitz, M., Eleftheriadis, A., and Wiegand, T., "H.264 hierarchical P coding in the context of ultra-low delay, low complexity applications", Picture Coding Symposium (PCS), Vol 1, pp. 146–149, 2010.
- [51] ISO/IEC-JTC1/SC29/WG11, "Call for Evidence on Multi-view Video Coding," October 2004.
- [52] D. Tian., P. Pandit., P. Yin., and C. Gomila., "Study of MVC coding tools", Joint Video Team (JVT) Doc. JVT-Y044, Shenzhen, China, Oct. 2007.
- [53] Philipp Merkle., Aljoscha Smolic., Karsten Müller. and Thomas Wiegand., "Efficient Prediction Structures for Multiview Video Coding.", IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Multi-view Video Coding and 3DTV, vol.17, pp.1461-1473, 2007.
- [54] H. Ozaktas and L. Onural, "Three-Dimensional Television: Capture, Transmission, Display", Springer, 2008.
- [55] MVC reference software obtained at, cvs-d: pserver:jvt-user@garcon.ient.rwth-aachen.de:/cvs/jvt checkout jmvc.
- [56] Gary J. Sullivan., and Thomas Wiegand., "Rate-distortion optimization for video compression", Signal Processing Magazine, IEEE, vol.15, no. 6, pp.74-90, 1998.
- [57] Thomas Wiegand., Michael Lightstone., Debargha Mukherjee., T. George Campbell., and Sanjit K. Mitra., "Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H. 263 standard." IEEE Transactions on Circuits and Systems for Video Technology, vol.6, no. 2, pp.182-190, April 1996.
- [58] VGA multi view sequences, <ftp://ftp.ne.jp/KDDI/multiview>
- [59] XGA multi view sequences, <http://www.tanimoto.nuee.nagoya-u.ac.jp/>
- [60] JM version 18.2 (<http://iphome.hhi.de/suehring/tml>)
- [61] Ye-Kui., M.M.Hannuksela., S.Pateux., A.Eleftheriadis., and Stephan wenger., "System and Transport Interface of SVC", IEEE Transactions on Circuits and Systems for Video Technology, vol.9, pp.1149-1163, September 2007.
- [62] G. Bjøntegaard "Calculation of average PSNR differences between RD-Curves", ITU-T SG16 Q.6, document VCEG-M33, 2001.

- [63] T. Wiegand., and B. Girod., "Multi-Frame Motion-Compensated Prediction for Video Transmission", Kluwer, 2001.
- [64] Y. Su., and M. Sun., "Fast multiple reference frame motion estimation for H.264/AVC," IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, No. 3, pp. 447-452, March 2006.
- [65] Li Bin., Jizheng Xu., Houqiang Li., and Feng Wu., "Optimized reference frame selection for video coding by cloud." in Proceedings of Multimedia Signal Processing (MMSP), 2011.
- [66] Q. Shen., Y.-K. Wang., M. M. Hannuksela., H. Li., and Y. Wang., "Buffer requirement analysis and reference picture management for temporal scalable video coding," Proceedings of International Packet Video Workshop, pp. 91-97, Nov. 2007.
- [67] H. Schwarz., D. Marpe., and T. Wiegand., "SVC core experiment 2.1: Inter-layer prediction of motion and residual data," ISO/IEC JTC 1/SC 29/WG 11, doc. M11043, Redmond, WA, USA, July 2004.
- [68] Segall, C. Andrew., and Gary J. Sullivan., Spatial scalability within the H. 264/AVC scalable video coding extension. IEEE Transactions on Circuits and Systems for Video Technology, vol.17, no.9, pp.1121-1135, September 2007.
- [69] François., J Vieron., and V Bottreau., "Interlaced Coding in SVC," IEEE Transactions on Circuits and Systems for Video Technology, vol.17, no.9, pp.1136-1148, September 2007.
- [70] Schwarz, Heiko., Detlev Marpe., and Thomas Wiegand., "Overview of the scalable video coding extension of the H. 264/AVC standard". IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 9, pp.1103-1120, September 2007.
- [71] Gary J. Sullivan., Ohm., Han W. J., and Wiegand, T., "Overview of the high efficiency video coding (HEVC) standard". IEEE Transactions on Circuits and Systems for Video Technology, vol.22, no.12, pp.1649-1668, December 2012.
- [72] JSVM reference software obtained at, cvs-d: pserver:jvt-user@garcon.ient.rwth-aachen.de:/cvs/jvt checkout jsvm.
- [73] S. Sun., and J. Reichel., AHG Report: "Spatial Scalability Resampling", Joint Video Team, JVTR006, 2006.
- [74] B. W. Micallef., and C. J. Debono., "An Analysis on the Effect of Transmission Errors in Real-time H.264-MVC Bit-streams," in Proceedings of Mediterranean Electrotechnical Conference (MELECON), Malta, April 2010.



- [75] T. Stockhammer., M. M. Hannuksela., and T. Wiegand., "H.264/AVC in Wireless Environments," IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 657-673, July 2003.
- [76] Y. Dhondt., and P. Lambert., "Flexible Macroblock Ordering, an Error Resilience Tool in H.264/AVC," 5th FTW PhD Symposium, Faculty of Engineering, Ghent University, 2004.
- [77] S. Kumar., L. Xu., M. K. Mandal., and S. Panchanathan., "Error Resiliency Schemes in H.264/AVC Standard," Journal of Visual Communication and Image Representation, Vol. 17, no. 2, pp. 425-450, April 2006.
- [78] C. Hewage., S. Warrall., S. Dogan., and A. M. Kondo., "Frame Concealment Algorithm for Stereoscopic Video Using Motion Vector Sharing," in Proceedings of IEEE International Conference on Multimedia and Expo (ICME), Germany, June 2008.
- [79] Y. Liu., J Wang., and H. Zhang., "Depth Image Based Temporal Error Concealment for 3D Video Transmission," IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, no. 4, pp. 600-604, April 2010.
- [80] C. L. Zitnick., S.B. Kang., M. Uyttendaele., S. Winderm., and R.Szeliski., "High-Quality Video View Interpolation using a Layered Representation," ACM SIGGRAPH and ACM transactions on Graphics, pp. 600-608, Los Angeles, USA, August 2004.
- [81] S. Liu., Y. Chen., Y.-K. Wang., M. Gabbouj., M. Hannuksela., and H. Li., "Frame Loss Error Concealment for Multiview Video Coding," in Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), pp. 3470-3473, USA, May 2008.
- [82] C. Bilen., A. Aksay., and G. Bozdagi Akar., "Two Novel Methods for Full Frame Loss Concealment in Stereo Video," in Proceedings of 26th Picture coding Symposium, Lisbon, Portugal, November 2007.
- [83] U. Fecker., J. Seiler., and A. Kaup., "4-D Frequency Selective Extrapolation for Error Concealment in Multi-View Video," in Proceedings of Multimedia Signal Processing (MMSP), Australia, October 2008.
- [84] S. Knorr., C. Clemens., M. Kunter., and T.Sikora., "Robust Concealment for Erroneous Block Bursts in Stereoscopic Images," in Proceedings on 3D Data Processing, Visualization and Transmission (3DPVT), September 2004.
- [85] C. Clemens., M.Kunter., S. Knorr., and T. Sikora., "A Hybrid Approach for Error Concealment in Stereoscopic Images," 5th International workshop Image Anal. Multimedia Interactive Services (WIAMIS), April 2004.

- [86] T. Chung., K. Song., and C.-S. Kim, "Error Concealment Techniques for Multi-view Video Sequences," *Advances in Multimedia Information Processing – PCM 2007*, pp-619-627, December 2007.
- [87] S. Tsekeridou., I. Pitas., "MPEG-2 Error Concealment Technique Based on Block-matching Principle," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 4, pp. 646-658, June 2000.
- [88] S. Argyropoulos., A. S. Tan., N. Thomos., E. Arikani., and M.G. Strintzis., "Robust Transmission of Multi-view Video Streams Using Flexible Macroblock Ordering and Systematic LT Codes," *3DTV CONFERENCE*, Kos, Greece, May 2007.
- [89] A. S. Tan., A. Aksay., G. B. Akar., and E. Arikani., "Rate-Distortion Optimization for Stereoscopic Video Streaming with Unequal Error Protection," *EURASIP Journal on Applied Signal Processing*, January 2009.

## **Biography**

### **Candidate Biography**

Palanivel Guruvareddiar received the Masters in Technology (M.Tech) degree in communications engineering from Vellore Institute of Technology, India, in 2005 with university second rank. He has over 10 years of industry experience in the field of video compression, perceptual computing, computer vision and advanced driver assistance systems. He has worked with clients from various geographies including USA, UK, Japan, Europe and Asia and worked in a number of projects from concept till productization. He is currently a specialist engineer in the advanced driver assistance systems technology group of transportation business unit, Tata Elxsi Limited, India. He is also working toward the Ph.D degree in the area of scalable multi view coding from Birla Institute of technology and science, Pilani, India. He has co-authored a number of publications in the field of video codecs and also filed a patent in the area of scalable video coding for multi view signal. He is a senior member of IEEE and also served as a reviewer in leading journals such as SPIE Journal of Electronic Imaging, Springer Journal on Signal, Image and video processing, Springer Journal on 3D research. His research interests include scalable video coding, 3D video compression and video communication systems.

### **Supervisor Biography**

Biju K. Joseph was born in Kerala, the southernmost state of India in 1966. He received B.Sc. and M.Sc. degrees in Physics from Mahatma Gandhi University in 1986 and 1988 respectively and the Ph.D. degree in Plasma Physics from Devi Ahilya Vishwa Vidyalaya, Indore for the research carried out at Institute for Plasma Research, Gandhinagar, India in 2002. From 1995 to 2001, he was a Research Scientist with Institute for Plasma Research. Since 2001 he is with Tata Elxsi Limited and has been actively involved in the development of Electronic circuit simulation, Digital Signal Processing algorithms for Multimedia Video codec, Digital Watermarking for Images, Video codec and hardware development for the real time H264 encoder and decoder. Currently his interest is in the development of Low Latency Long Range Wireless communication systems. He has co-authored a number of publications in the field of video codecs and also filed couple of patents in the area of scalable video coding for multi-view signal and multi-relay wireless communications.