LOW BITRATE MULTI-VIEW VIDEO CODING BASED ON H.264/AVC

HANY HANAFY MAHMOUD SAID

A thesis submitted in partial fulfilment of the requirement of Staffordshire University for the degree of Doctor of Philosophy

December 2015

ABSTRACT

Multi-view Video Coding (MVC) is vital for low bitrate applications that have constraints in bandwidth, battery capacity and memory size. Symmetric and mixed spatial-resolution coding approaches are addressed in this thesis, where Prediction Architecture (PA) is investigated using block matching statistics. Impact of camera separation is studied for symmetric coding to define a criterion for the best usage of MVC. Visual enhancement is studied for mixed spatial-resolution coding to improve visual quality for the interpolated frames by utilising the information derived from disparity compensation.

In the context of symmetric coding investigations, camera separation cannot be used as a sufficient criterion to select suitable coding solution for a given video. Prediction architectures are proposed, where MVC that uses these architectures have higher coding performance than the corresponding codec that deploys a set of other prediction architectures, where the coding gain is up to 2.3 dB. An Adaptive Reference Frame Ordering (ARFO) algorithm is proposed that saves up to 6.2% in bits compared to static reference frame ordering when coding sequence that contains hard scene changes.

In the case of mixed spatial-resolution coding investigations, a new PA is proposed that is able to save bitrate by 13.1 Kbps compared to the corresponding codec that uses the extended architecture based on 3D-digital multimedia. The codec that uses hierarchical B-picture PA has higher coding efficiency than the corresponding codec that employs the proposed PA, where the bitrate saving is 24.9 Kbps. The ARFO algorithm has been integrated with the proposed PA where it saves bitrates by up to 35.4 Kbps compared to corresponding codec that uses other prediction architectures. Visual enhancement algorithm is proposed and integrated within the presented PA. It provides highest quality improvement for the interpolated frames where coding gain is up to 0.9 dB compared to the corresponding frames that are coded by other prediction architectures.

ACKNOWLEDGEMENT

I would like to thank Allah, my lord for all his great blessing in my life.

Secondly, I would like to express my gratitude to my principle supervisor Prof. Mansour Moniri for his patient guidance and scientific support during my research study. I appreciate his encouragement throughout my PhD study.

I would like to thank my second supervisor Prof. Claude Chibelushi for his valuable inputs and encouragement.

I would like to thank Dr. Akbar Sheikh Akbari for his supervision in the first phase of my PhD study. Also, I would like to acknowledge Mohsin Abbas Malik for his effort during conducting the simulation experiments for impact of camera separation and stereoscopic video coding using block matching statistics.

I would like to thank Staffordshire University for providing me the opportunity to undertake PhD programme in addition to providing the resources that are needed to conduct the research. I would like to thank my colleagues for their support during the research study.

I would like to express my deepest gratitude to my deceased mother, without her continuous encouragement I would never have been able to achieve my targets. I would like to thank my beloved father, who gave me best friendship ever and would like to thank my wife and my daughter for their support.

This thesis is dedicated to the memory of my parents.

Hany Hanafy Mahmoud Said

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

3D-DMB	3D-Digital Multimedia Broadcast
3DTV	Three-Dimensional TV
AMRSC	Advanced Mixed-Resolution Stereo Coding
ARFO	Adaptive Reference Frame Ordering
AVC	Advanced Video Coding
BM	Block Matching
CBP	Coded Block Pattern
CfP	Call for Proposal
CIF	Common Intermediate Format
DCP	Disparity Compensation Prediction
DCT	Discrete Cosine Transform
DCVF	Disparity Compensation View Filtering
DE	Disparity Estimation
DIBR	Depth Image Based Rendering algorithm
DIP	Diagonal Inter-view Prediction
DPB	Decoded Picture Buffer
DSCQS	Double-Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
DV	Disparity Vector
DVB	Digital Video Broadcasting
EPSNR	Edge Peak Signal to Noise Ratio
FIR	Finite Impulse Response
FPS	Frames Per Second
FR	Full spatial-Resolution
FTV	Free viewpoint TV
GDV	Global Disparity Vector
GoGOP	Group of GOP
GOP	Group Of Pictures
GSM	Global System for Mobile communications
HBP	Hierarchical B-Picture
HDTV	High Definition TV
HEVC	High Efficiency Video Coding

HVS	Human Visual System
IC	Illumination Compensation
IDR	Instantaneous Decoder Refresh
ITU	International Telecommunication Union
IVP	Inter-View Prediction
JVT	Joint Video Team
Kb/s	Kilobits per second
LDV	Layered Depth Video
LR	Low spatial-Resolution
MAD	Mean Absolute Difference
MB	Macroblock
MC	Motion Compensation
MCTF	Motion Compensation Temporal Filtering
ME	Motion Estimation
MOP	Matrix Of Picture
MPEG	Motion Picture Expert Group
MS	Motion Skip
MSE	Mean Square Error
MSSIM	Mean Structural SIMilarity
MV	Motion Vector
MVC	Multi-view Video Coding
MVD	Multi-view Video plus Depth
MVI	Multi-View Imaging
MVV	Multi-View Video
NAL	Network Abstraction Layer
NIP	Normal Inter-view Prediction
PA	Prediction Architecture
PPS	Picture Parameter Set
PSNR	Peak Signal to Noise Ratio
PSTN	Public Switched Telephone Network
QP	Quantization Parameter
QVGA	Quarter Video Graphic Array
R-D	Rate-Distortion
REF	REference Frame

RFO	Reference Frame Ordering
RFS	Reference Frame Selection
SAD	Sum of Absolute Difference
SATD	Sum of Absolute Transformed Difference
SD	Standard Definition
SEI	Supplemental Enhancement Information
SI	Spatial Index
SNR	Signal-to-Noise Ratio
SPS	Sequence Parameter Set
SSE	Sum of Square Error
SSIM	Structural SIMilarity index
StSD	Stereoscopic Structural Distortion
SVC	Scalable Video Coding
SVPS	Sequential View Prediction Structure
TI	Temporal Index
TS	Transport Layer
UP	UP-sample
VCEG	Video Coding Experts Group
VCL	Video Coding Layer
VE	Visual Enhancement
VGA	Video Graphic Array
VLSI	Very Large Scale Integration
VPD	Video Plus Depth
VQM	Video Quality Metric
VSP	View Synthesis Prediction
XGA	eXtended Graphics Array

CHAPTER 1. INTRODUCTION

Multi-view video is an exciting technology which has great applications in our life. The first section will introduce multi-view video, and set of low bitrate applications are then outlined. The multi-view video chain is illustrated, where the coding component is highlighted. Multi-view video codec standards followed by research problem and motivations are also presented. The aim and objectives of this research are then outlined, followed by thesis structure at the end of this chapter.

1.1 Multi-view video overview

Multi-view video is a set of videos which are captured using synchronised cameras that are closely located at different viewpoints. The majority of videos in the set contain similar visual information, where the variance is due to disparity, occlusion and illumination effects (Jeon et al., 2009; Dufaux et al., 2013). The number of videos embedded in multi-view video is greater than one. Stereoscopic video is a special case of multi-view video that is inspired from the Human Visual System (HVS), where each eye perceives the corresponding video (Vetro et al., 2011). Multi-view video facilitates 3D perception by supporting a set of cues. These cues are (Dodgson, 2005; Boev et al., 2011a):

- Stereo parallax: each human eye receives a slightly different image.
- Motion parallax: provides perceptual cues about the change in motion, distance linked by depth perception.
- Accommodation: the ability to see sharply all objects at various distances.
- Ocular convergence: the human focuses on a certain object, both eyes move inward to get a single binocular vision.

Multi-view video would be used either to enrich the user experience through watching the scene from different viewpoints such as free-viewpoint Television (FTV) or viewing two slightly different views concurrently such as three-dimension Television (3DTV) (Tanimoto, 2009; Lee et al., 2010; Smolic, 2011). The former scenario extends view navigation functionality while the latter scenario stimulates depth perception for HVS. Multi-view video has opened a wide range of applications which include; entertainment, immersive teleconferencing, facial recognition and many other exciting applications (Minoli, 2011; Dufaux et al., 2013).

Since this thesis focuses on enabling multi-view video for applications that prefer low bitrates, the next section will outline a set of these applications for transmission scenario.

1.2 Low bitrate applications

Although portable devices such as tablets and smart phones support high bitrate, they have limited energy resources (battery) (Miao et al., 2009). Since smart phones host many applications, the power consumption is increasing rapidly in comparison to battery capability (Miao et al., 2009). Power consumption increases when decoding videos, that is affected by frame rate, spatial-resolution and bitrate (Lin et al., 2007). Video conferencing requires efficient bandwidth utilisation, low processing delay and better video quality, where spatial-resolution is usually Common Intermediate Format (Schwarz et al., 2007). It could be deployed using a wireless cameras array over IEEE 802.11b network (Yang & Goodwin, 2005). Telemedicine provides medical services such as remote diagnosis of patients ailments and remote surgery to patients who live in rural areas that lack on-site medical facility (Paul & Sorwar, 2007). Low bitrate is preferable for this type of application in order to reduce cost of medical service (Hewage et al., 2013). Video telephony is usually deployed using low bitrate transmission via Public Switched Telephone Network (PSTN) and Global System for Mobile communications (GSM) (Eisert, 2000; Kwon & Driessen, 2001). Internet Protocol Television is one of the entertainment applications. It provides low quality video that is coded and transmitted to end-user, where received signal is decoded and post processed to remove strong artefacts (Shao et al., 2009). In surveillance, wireless ZigBee networks are used to enable monitoring activities for short time periods using low bitrate transmission via IEEE 802.15.4 protocol (Zainaldin et al., 2008).

This section has outlined some applications that prefer low bitrate in transmission scenario. Generally, low bitrate is a constraint that is enforced in applications, where end-users' devices have limited bandwidth connections, portable devices with limited battery capability, restricted amount of memory or small displays. The next section will introduce the multi-view video chain that begins from capturing to display, where each component is briefly outlined.

1.3 Multi-view video chain

The multi-view video chain is depicted in Figure 1-1, where each block represents a component within the chain. Since this thesis focuses on coding multi-view video, the main component has been highlighted by grey colour.

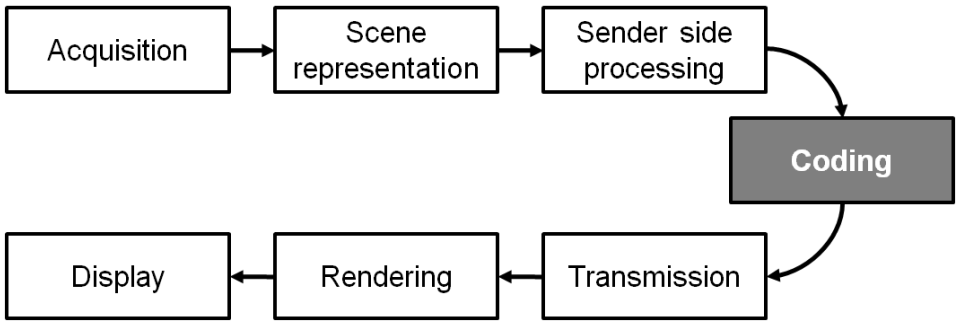


Figure 1-1 Multi-view video chain

Acquisition: involves capturing a scene from a set of cameras. The most widely used cameras' arrangements are (Lee et al., 2010):

- 1D parallel (linear): cameras are positioned as an array either in vertical or horizontal direction. Most applications prefer horizontal direction to be consistent with motion parallax for HVS.
- 2D parallel: cameras are placed in horizontal and vertical directions.
- 1D arc (coplanar): cameras are positioned in convergent setup toward the scene centre. The videos are usually rectified in order to easily locate the corresponding points among views in the horizontal direction.

Scene representation: the following are different formats for multi-view video, they are (Alatan et al., 2007; Morvan et al., 2008; Smolic, 2011):

- Texture: All views are represented by their texture videos.
- Depth: Part of visual data is texture alongside its depth map. It represents the pixel distances from the camera using grey scale image/video.
- Model: video is either represented by foreground and background objects or through 3D meshes with their texture mapping.

Sender side processing: It includes colour correction; white balancing, finding camera parameters; rectifying convergence multi-view video, objects segmentation and depth estimation (Smolic, 2011).

Coding: Multi-view video is a superset of monoscopic video. Thus the main obstacle for transmission or storage is the huge size of multi-view video (Chen et al., 2009b; Vetro, 2010), e.g. the required bandwidth for transmitting 2D video with Video

Graphics Array (VGA) resolution size using three full colour sampling with 30 Frames Per Second (FPS) will be 26.37 Mb/s. When extending this video to involve eight cameras, the size of its raw data will be 210.94 Mb/s. This example shows the high bitrate needed for transmitting multi-view video. Coding is the direct solution for multi-view video as it offers practical solution when this type of video is transmitted or stored. Therefore visual data is coded first in order to reduce its size without significantly degrading the visual quality of original data. Thus coding multi-view video is usually considered in any multi-view application.

A video codec exploits different types of visual information redundancies using hybrid video coding. It aims to minimise the number of bits required to represent the visual data. Therefore, the efficiency of video coding would be measured either by number of kilobits per second or number of bits per pixel for transmission and storage respectively. A video codec has two entities: an encoder and a decoder. The former compresses visual data and the latter decompresses the coded video prior to display. A video codec would use one of the following coding formats:

- Texture (image-based): uses only texture views.
- Depth: includes three subcategories
  - Video plus depth: uses one texture and one depth. Depth map supports view synthesis for narrow scene navigation. The rendered image would suffer from disocclusion which affects image quality.
  - Layered depth video: involves two texture views in addition to the depth and occlusion layers. It resolves the challenge of occlusion more efficiently at the expense of more complexity than video plus depth. The rendered image is affected by shadow and reflection area.
  - Multi-view plus depth: uses *N* texture views with their corresponding depth maps. It supports wider range of scene navigation at the expense of massive amount of data with respect to previous formats.
- Model: supports free view-point navigation. Automatic segmentation and high quality animated objects are the main challenges for this format.

The texture-based coding is the simplest format because the sender needs neither to estimate object depth nor model them, while the transmitted videos will be displayed directly at the decoder side.

Transmission: This is the carrier media, where the multi-view video would be spread using the broadband and broadcast connections (e.g. digital video broadcast) (Mignone et al., 2011). It is broadcast either via frame-compatible or service-compatible formats (Vetro et al., 2011; Dufaux et al., 2013).

- Frame-compatible: supports stereoscopic video, where both views are spatially multiplexed prior to coding. The most common arrangements are side-by-side and top-and-bottom.
- Service-compatible: provides more flexibility than the previous format, where a legacy decoder would be able to extract a single view. Stereoscopic video is supported via multi-view video coding standards.

Rendering: generates novel views using perceived coded views. Each scene representation format has its rendering characteristics (Stoykova et al., 2007; Smolic, 2011; Tanimoto, 2012).

- Texture: interpolates views with limited quality.
- Depth: synthesises views using a Depth Image Based Rendering algorithm (DIBR). They are obtained by projecting the pixel from image plane into 3D space then back projecting it into different camera plane.
- Model: A new view is constructed after obtaining a visual hull followed by surface extraction, surface smoothing and mesh complexity reduction.

Display: includes stereoscopic, auto-stereoscopic, volumetric and holographic displays (Benzie et al., 2007; Smolic, 2011).

- Stereoscopic: shows single binocular disparity. It requires eye-glasses to filter different images to the corresponding eye.
- Auto-stereoscopic: supports multiple binocular disparities via motion parallax. There are two cases for displaying multi-view video:
  - Two-views: integrates head-tracking system which identifies the head's position, thus the videos are displayed correctly towards the user.
  - Multi-view imagery: uses either lenticular sheets or parallax barrier to distribute several images to set of viewing zones. These zones are the valid areas, where user would get 3D perception.
- Volumetric (multi-planar): image is displayed within volume of space, where each point of light has corresponding point in 3D space that entails massive size of video. Although it supports viewing 3D from wide range of viewpoints, it has several challenges such as capturing real scenes.
- Holographic: uses a photographic plate to reconstruct the objects using laser projection, diffraction and interference. It supports high quality images with depth cues. Due to the current very large scale integration technology, this type of display is usually found at research centres.

Since this thesis targets multi-view video coding, the following section will briefly categorise multi-view video codec standards.

1.4 Multi-view video codecs taxonomy

Simulcast video coding is most straightforward solution for coding multi-view video since it compresses each view separately. It can be used by any monoscopic video codecs, e.g. H.264/AVC or HEVC. Multi-view video codec standards compress jointly the given multi-view video, where spatial redundancies among neighbouring views are exploited. These multi-view video codec standards are:

- MPEG-2 Multi-view profile which was finalised in 1996 targeting stereoscopic videos. It encodes independently the left view while the right view uses the previous decoded right frames in addition to the neighbouring left frames in prediction (Chen & Luthra, 1997; Ohm, 1999).
- MPEG-C Part 3 was standardised in 2007. It is based on video plus depth, where each layer is coded separately. The average increase in bitrate is 8% extra with respect to 2D video compression (Bourge et al., 2006).
- H.264/MVC was released in 2008. It provides efficient coding for multi-view video through extending the motion estimation of H.264/Advanced Video Coding (AVC) (Chen et al., 2009b). This codec is deployed in entertainment applications, e.g. Blu-ray 3D Discs. It contains stereoscopic video which includes a base view and dependent view. Stereoscopic video would be decoded either by 3D Blu-ray player that provides 3D video or via 2D Blu-ray player that extracts a base view only and ignores the other view (Vetro et al., 2011; Tanimoto, 2012; Dufaux et al., 2013).
- MVC+D supports coding texture views and depth maps, where both are coded independently. H.264/MVC decoder is therefore capable of extracting and decoding the coded texture views (Hannuksela et al., 2013). Although this codec applies few changes to encapsulate coded texture views and depth maps into single stream, it does not exploit the redundancy that exist among texture views and their depth maps (Chen & Vetro, 2014).
- 3D-AVC applies changes at macroblock level to supports additional coding tools that exploit the redundancies among texture views and their depth maps (Zhang et al., 2013; Hannuksela et al., 2013). Part of the coding tools are neighbouring block-based disparity vector derivation, inter-view motion prediction and view synthesis prediction (Chen & Vetro, 2014). 3D-AVC is more coding efficient than MVC+D, where the bitrate reduction is on average 14% (Hanhart et al., 2014).
- MV-HEVC enables coding the texture views using HEVC, where few changes are applied to enable inter-view prediction among neighbouring views (two and three views are supported) (Yuan et al., 2015). These changes include allowing

reference frames that belong to the base view to be used during predicting frames that belong to dependent views (Aflaki et al., 2014; Sansli et al., 2014).

- 3D-HEVC supports MVD coding format, where each video is associated with the corresponding depth map. Each texture frame will be followed by its depth map and interleaved with the successive views. It is based on High Efficiency Video Codec (HEVC) in order to code high and ultra-high definition video resolution efficiently. This codec supports inter-view motion prediction, residual prediction and view synthesis prediction (Tech et al., 2015). This extension acts as the starting phase in standardising 3D video coding based on HEVC, where the first report for this extension was published in 2012 (Tech, 2012). 3D-HEVC is a suitable alternative to H.264/MVC when targeting autostereoscopic displays, where a subset of the texture videos with their depth maps will be transmitted. At the receiver side, the corresponding views will be decoded and displayed in addition to rendering novel views using DIBR algorithm (Müller et al., 2013).

The MPEG-2 Multi-view profile has limited usage due to the challenges of display and hardware capabilities at that time (Smolic et al., 2007). Disocclusion is the main challenge for MPEG-C Part 3 standard that affects the video quality for the rendered views, where the occluded areas in the main texture video cannot be rendered efficiently (Vetro, 2010). H.264/MVC provides efficient coding solution for multi-view video at the expense of demanding huge computational complexity and large memory requirements compared to simulcast video coding (Zhang et al., 2008). It puts a limitation for predicting frames using only temporal and spatial frames, on the contrary to H.264/AVC that provides full flexibility in determining the frames that are included in the prediction through multi-reference prediction coding tool (Chen et al., 2009b). Since MVC+D, 3D-AVC and 3D-HEVC support coding MVV with their associate depth maps, the rendered views are affected by the quality of depth map and the amount of disocclusion areas. MV-HEVC is based on the recent video codec HEVC. Although HEVC provides better coding efficiency than H.264/AVC (bitrate reduction is in the range of 50%), its encoder consumes a higher coding time than H.264/AVC encoder by at least a factor of four (Sullivan et al., 2012; Bossen et al., 2012). It uses 16-bit data format instead of 8-bit as in H.264/AVC, therefore HEVC requires more memory bandwidth than H.264/AVC (Bossen et al., 2012). HEVC is suitable for applications that target ultra-high definition displays and parallel processing capabilities (Ahn et al., 2014). H.264/AVC remains a powerful coding standard for low bitrate applications that provides full flexibility for inter-picture

prediction compared to H.264/MVC in addition to less computational complexity and memory bandwidth than HEVC.

1.5 Research problem and motivations

Multi-view video has a wide area of applications nowadays, as it provides the user either the ability to watch multiple views simultaneously or viewing the scene from different viewpoints. These advantages entail higher amount of visual data than 2D video that is proportional to the number of cameras used in capturing the scene. Therefore coding multi-view video is an inevitable stage in applications that support multi-view video. Low bitrate in the context of multi-view video coding refers to coding the given multi-view video at lowest acceptable quality. This is defined for each multi-view video using either quantisation parameter or bitrate. According to common test conditions for MVC, average low bitrate per camera is in range of 128 Kbps to 768 Kbps (Su et al., 2006). This constraint is desirable when end-users devices have limited bandwidth connections, portable devices with limited battery capability, restricted amount of memory or small displays. Low bitrate multi-view video coding targets the coding performance in terms of rate-distortion without significantly increase the computational complexity and memory consumption of MVC. H.264/AVC is an essential coding standard for low bitrate applications that provides more flexibility for inter-picture prediction than H.246/MVC and it requires less computational complexity and memory bandwidth than HEVC.

The general block diagram for coding multi-view video using H.264/AVC is depicted in Figure 1-2 that enables transmitting multi-view video between sender and receiver sides. The captured videos from neighbouring cameras are multiplexed into single sequence. During compression, each frame is divided into blocks. Each block is predicted, where the residual signal is transformed and quantised. Entropy coder compresses the prediction information and transformed coefficients, prior to transmission. The received video is decoded and de-multiplexed prior to display. A multi-view video codec focuses on exploiting visual redundancies among neighbouring views (inter-view correlation), using prediction (highlighted by a grey colour in Figure 1-2). This justifies the importance of the prediction component in coding multi-view video rather than simulcast video coding that compresses each view separately (Vetro et al., 2011).

A resolution-based coding approach is an attractive solution when addressing multi-view video coding at low bitrates. It requires neither depth-map nor segmentation. It avoids the challenges related to depth estimation, holes filling and

automatic segmentation. Symmetric multi-view video coding has shown superior coding performance at low bitrates when it is compared to other coding approaches (Strohmeier & Tech, 2010; Saygili et al., 2010; De Silva et al., 2013). Symmetric multi-view video codec increases the quantisation parameter in order to meet the target bitrate. On the other hand, mixed spatial-resolution multi-view video coding reduces the amount of visual data, where the total perceived quality is close to the quality of full spatial-resolution frames due to the suppression theory (Aflaki et al., 2013a). According to this theory, high frequency components that exist in the full spatial-resolution frames compensate the corresponding components in the lower spatial-resolution frames (De Silva et al., 2012). The mixed spatial-resolution multi-view video coding approach reduces coding complexity and improves objectively the coding performance compared to symmetric coding (Fehn et al., 2007; Brust et al., 2009; Aflaki et al., 2013a). Therefore, symmetric MVC and mixed spatial-resolution MVC are used in the investigations reported in the thesis.

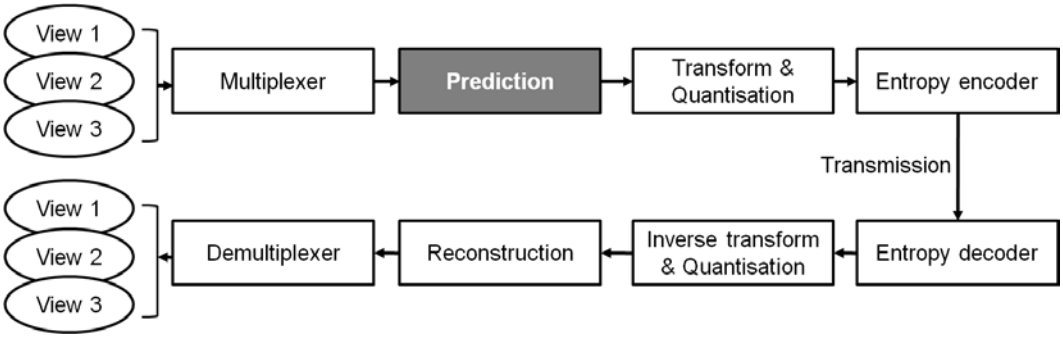


Figure 1-2 General block diagram for H.264/AVC based multi-view video coding

Symmetric multi-view video coding is beneficial for multi-view video that contains a significant amount of spatial redundancies among neighbouring views. When the amount of these redundancies is insignificant, then simulcast video coding should be used since it requires less computational complexity than Multi-view Video Coding (MVC). The amount of spatial redundancies depends on camera separation, where closer cameras have higher degree of visual correlation than sparse cameras. Although several studies addressed the effect of camera separation on coding performance of MVC, still the criterion for the best usage of multi-view video coding is not defined (Merkle et al., 2007a; Bouyagoub et al., 2010). The prediction architecture is a main part in the prediction component of H.264/AVC since it defines the reference frames that are used in the prediction (reference frame selection) alongside defining how to address these frames during compression (reference frame ordering). Although a typical prediction architecture of H.264/MVC achieves

efficient coding gain compared to other prediction architectures, it has significant computational complexity and memory requirements (Zhang et al., 2008). Several prediction architectures have been proposed in the literature. Parts of these architectures justify neither reference frame selection nor reference frame ordering (Oka et al., 2004; Fecker & Kaup, 2005; Oh & Ho, 2007; Flierl et al., 2007). A few studies looked into the statistical analysis of block matching as a reliable technique to derive prediction architectures (Kaup & Fecker, 2006; Merkle et al., 2007a). Since they do not deploy all coding tools of H.264/AVC, the efficiency of inter-picture prediction is degraded. Still there are no clear clues about reference frame selection that should be used when H.264/AVC operates at low bitrates. Although a few studies have proposed different mechanisms for reference frame reordering, they do not provide a practical solution that fits the requirements of low bitrate applications (Pourazad et al., 2009a; Seungwook & Yang, 2011).

In the context of mixed spatial-resolution multi-view video coding, frames that belong to neighbouring views may have different spatial-resolution. Therefore, the reference frames need further processing (either decimation or interpolation) before deploying inter-view prediction. Although the effect of deploying inter-view prediction direction among mixed spatial-resolution frames is addressed in the literature, the outcomes might be influenced by asymmetric quality settings at the point of conducting experiments (Brust et al., 2010). Therefore, there is a need to investigate different inter-view prediction directions to reveal the challenges when coding mixed spatial-resolution multi-view video. There are different decimation and interpolation methods, where there is no clear efficient method in terms of coding gain and computational complexity (Aksay et al., 2006; Fehn et al., 2007; Aflaki et al., 2013b). It is important to deploy suitable methods for decimation and interpolation to enable inter-view prediction without significantly increasing the computational complexity. Although a few prediction architectures are deployed, they are either inherited from a typical prediction architecture of H.264/MVC or there is no theoretical justification behind reference frame selection (Fehn et al., 2007; Chen et al., 2009a). Mixed spatial-resolution multi-view video could be used for free-viewpoint video (Garcia et al., 2010b). On the other hand, eye fatigue has been reported when viewing coded mixed spatial-resolution stereoscopic video (Jain et al., 2012, 2014). This entails that blurring artefacts for the asymmetric spatial-resolution coding approach should not be entirely ignored when it is deployed in the context of multi-view video coding. Several algorithms have looked into enhancing the visual quality of the interpolated frames (coded low spatial-resolution frames) through reducing the amount of blurriness at the receiver side (Tech et al., 2009a; Najafi, 2012). Since these studies

did not provide an efficient low complexity algorithm, an efficient solution is needed to enhance the visual quality for these frames.

1.6 Aim and objectives

The aim of this research is to investigate multi-view video coding using H.264/AVC when the codec operates at low bitrate.

The following list summarises the research objectives:

- Conduct a literature survey on:
  - Coding approaches that are used for low bitrate video codecs.
  - Multi-view video coding, particularly its prediction architectures.
- In the context of symmetric spatial-resolution multi-view video coding:
  - Determine the impact of camera separation on the coding performance of multi-view video codec.
  - Investigate prediction architectures using a statistical analysis of block matching among different reference frames.
  - Investigate reference frame reordering.
- In the context of mixed spatial-resolution multi-view video coding:
  - Explore the effect of deploying inter-view prediction using full and low spatial-resolution reference frames.
  - Investigate suitable methods for decimating and interpolating reference frames.
  - Investigate prediction architectures using block matching statistics for both; full and low spatial-resolution frames.
  - Enhance visual quality for the coded low spatial-resolution frames.

1.7 Thesis structure

The thesis contains six chapters; it is organised as follows

- **Chapter 1** introduces multi-view video including low bitrate applications and multi-view video chain. Multi-view video codec standards are then outlined. The research problem and motivations are defined. Aim and objectives of the investigations are then presented, followed by the structure of the thesis that is listed at the end of this chapter.
- **Chapter 2** presents the background to H.264/AVC key technologies. Parts of its coding tools are highlighted that are relevant to the thesis. H.264/MVC and its

limitations are outlined. Video quality metrics are then categorised, where objective quality metrics that are used in the thesis are outlined.

- **Chapter 3** categorises low bitrate multi-view video codecs. Block matching efficiency and prediction architectures are reviewed in the context of symmetric multi-view video coding, while prediction architectures and visual enhancement algorithms are reviewed in the context of mixed spatial-resolution MVC. A review summary and list of research investigations are then presented.
- **Chapter 4** focuses on symmetric multi-view video coding, where the impact of camera separation on the coding performance of MVC is studied. Prediction architectures are investigated for stereoscopic and multi-view video coding followed by examining reference frame reordering.
- **Chapter 5** targets asymmetric spatial-resolution multi-view video coding. It presents the effect of deploying different inter-view prediction directions on the coding performance of the MVC. It then examines different methods for decimating and interpolating reference frames. Prediction architectures are then explored via block matching statistics. The feasibility of reducing blurriness in the interpolated frames is investigated.
- **Chapter 6** starts with the conclusions of research outcomes. A set of potential studies are then outlined that would provide future research directions.

CHAPTER 2. BACKGROUND

This chapter presents the background that is relevant to H.264/AVC, H.264/MVC and video quality metrics. It starts with H.264/AVC and its coding tools, where parts of these tools that are relevant to the research investigations are highlighted. H.264/MVC is presented, where its typical prediction architecture and its limitations are demonstrated. Video quality metrics are outlined, where the metrics used in this research are illustrated.

2.1 H.264/AVC standard

This section focuses on the key technologies that provide coding efficiency behind H.264/AVC. This coding standard (named as MPEG-4 part 10) has been developed via Joint Video Team (JVT). It reflects a joint collaboration between Video Coding Experts Group and Moving Picture Experts Group (MPEG) committees. Parts of codec applications are: mobile TV, HD broadcasting and video conference.

H.264/AVC is a hybrid video codec that relies on prediction and transformation (Marpe et al., 2005; Richardson, 2010). Figure 2-1 depicts H.264/AVC block diagram, where T, Q, T^{-1}, Q^{-1}, ME, MC and DPB are transform, quantisation, inverse transform, inverse quantisation, motion estimation, motion compensation and Decoded Picture Buffer respectively. The red dashed box defines the relevant parts of the prediction component that includes ME, MC, intra-prediction, DPB, List buffers and the Deblocking filter. The blue dashed box identifies the relevant parts of prediction architecture for H.264/AVC that involves DPB and List buffers. The input frame is divided into macroblocks; each has 16×16 pixels. Each macroblock is predicted, where the residual signal[1] is transformed and quantised. Entropy coder compresses control data, prediction information and transformed coefficients. Since H.264/AVC codec uses coded pictures to deploy inter-picture prediction for next frames, it needs to decompress the coded frame. It applies an inverse operation for both quantisation and transformation, and then it deploys motion compensation to reconstruct the frame following by Deblocking filtering. The resultant frame is stored in DPB to guarantee identical reference frame at encoder and decoder sides. The next subsections outline the functionality of each block.

[1] It is the resulted signal after subtracting prediction from original block.

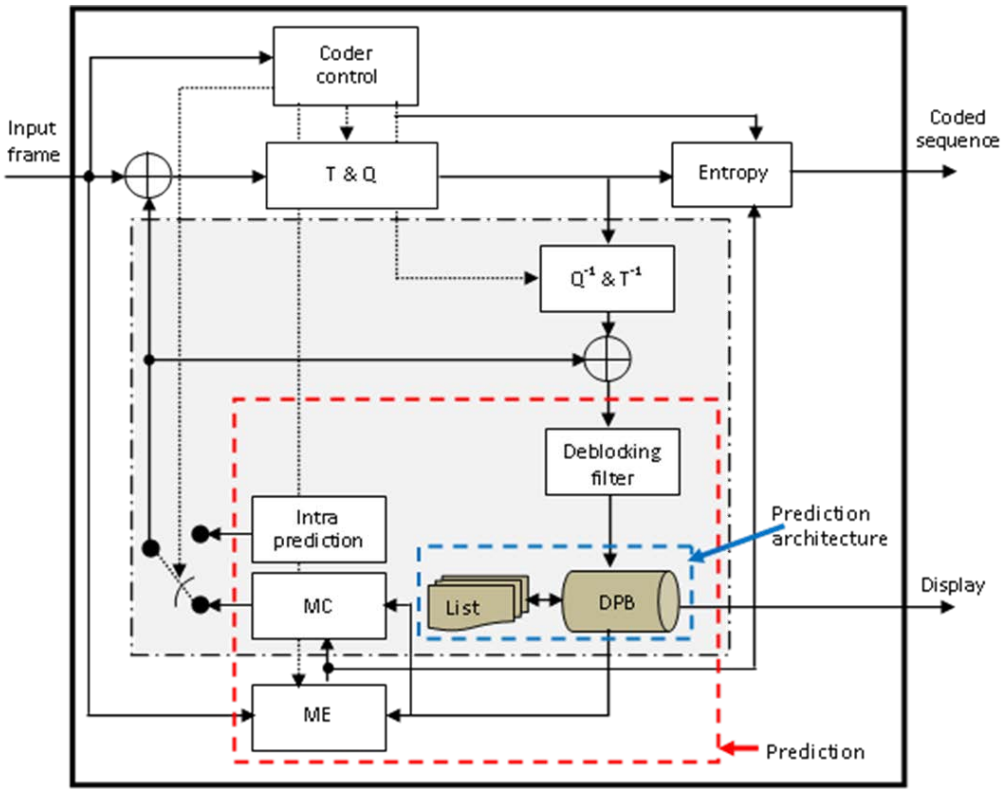


Figure 2-1 H.264/AVC block diagram for monoscopic video coding (Schwarz et al., 2006; Richardson, 2010)

2.1.1 Prediction

Prediction is core component for H.264/AVC. The next subsection outlines different prediction, macroblock and frame types.

2.1.1.1 Prediction, macroblock and frame types

There are two types of prediction:

- Intra-prediction exploits the spatial redundancy for a macroblock using its neighbouring macroblocks. Since video could be segmented into foreground and background objects, neighbouring pixels among these objects share significant spatial correlations which would be exploited by intra-prediction.
- Inter-picture prediction exploits temporal redundancy. Successive frames contain foreground objects in addition to static background. Therefore, there are significant temporal correlations among neighbouring temporal frames. Predictive coding exploits these redundancies as the new changes in the frame are coded instead of coding the entire frame through macroblocks that belong to reference frames (previously coded frames).

The macroblock is the basic unit in the codec that is either predicted by intra-predication or inter-picture prediction as depicted in Figure 2-2. There are three types of macroblocks (Nukhet & Tunali, 2005; Marpe et al., 2006b; Richardson, 2010)

- I-macroblock: allows intra-prediction using 4×4, 8×8 or 16×16 luma prediction. Luma prediction using 16×16 has four modes, they are horizontal, vertical, DC and planar. Luma predictions using 4×4 and 8×8 choose one from nine modes. Mode 0 to eight extrapolate samples in vertical, horizontal, DC, diagonal down-left, diagonal down-right, vertical-left, horizontal-down, vertical-right and horizontal-up directions respectively.
- P-macroblock: uses inter-picture prediction, where the samples that belong to reference frames are stored in DPB. Forward reference frames are used to get best block matching for current P-macroblock.
- B-macroblock: expands the capability of inter-picture prediction deployed in P-macroblock. It allows bi-prediction using forward and backward reference frames. Therefore, macroblock could be predicted either by forward, backward or bi-predictions. This entails more memory is needed than P-macroblock in order to store forward and backward reference frames.

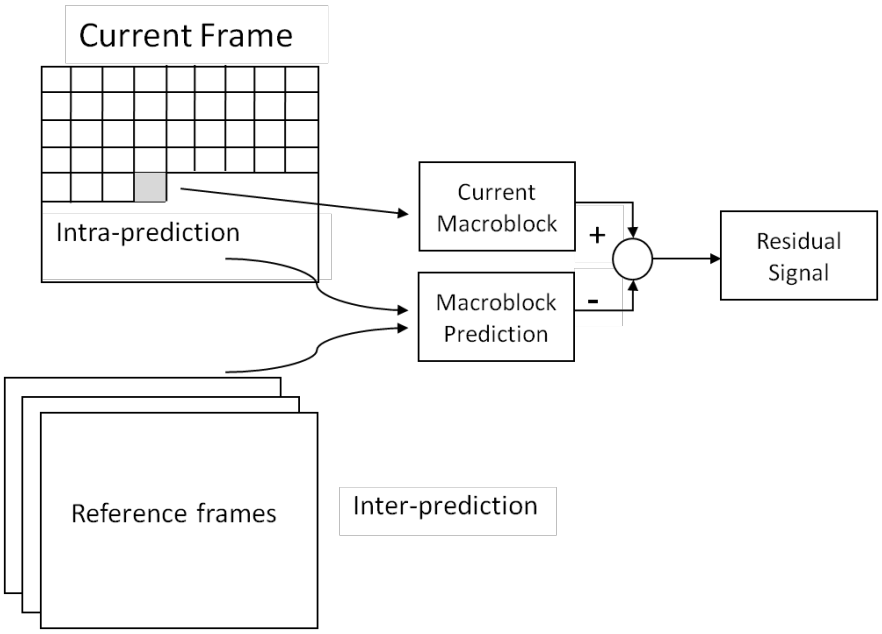


Figure 2-2 Prediction types of H.264/AVC (Richardson, 2010)

There are three types of frames that are supported in H.264; they are I-frame, P-frame and B-frames. I-frame contains only I-macroblocks while P-frame has combinations of I-macroblocks and P-macroblocks. B-frame contains all types of macroblocks. Although I-frame provides less coding efficiency than P-frame and B-

frame, it does not deploy motion estimation. This enables fast coding in addition to supporting temporal random access. Inter-picture prediction could deploy weighted prediction, where the predicted block is weighted using its temporal distance to current frame. Both P-frame and B-frame could be used as source for inter-picture prediction. B-frame gets efficient coding, where bi-prediction gives more accuracy than uni-prediction which is deployed in P-frame (Wong et al., 2011).

2.1.1.2 Coding tools

The codec supports a set of coding tools; they are multi-reference prediction, coding modes, sub-pixel ME and MC in addition to Deblocking filter.

Multi-reference prediction is a key feature behind inter-picture prediction that provides H.264/AVC with the flexibility in selecting suitable source among available reference frames. Motion estimation searches for best block matching within temporal reference frames that are already decoded and stored in DPB. The chosen reference frame is signalled through transmitting its index in List buffer. This buffer stores the reference frames indices. List 0 is dedicated for forward prediction during coding P-frame, while List 0 and List 1 are used to store forward and backward reference frames' indices respectively when B-frame is coded. Multi-reference prediction enables predicting macroblock using multiple reference frames, where each macroblock could be predicted using single or multiple reference frames. This flexibility provides the codec the capability to provide accurate prediction for the macroblock at the expense of computational complexity that is linked to the number of reference frames in addition to increasing amount of memory to store the relevant reference frames (Richardson, 2010).

The prediction uses different coding modes (block sizes) as depicted in Figure 2-3. Macroblock could be divided by 16×16, 16×8, 8×16 or 8×8 macroblock partitions. Therefore a macroblock could contain one 16×16 partition, two 8×16 partitions, two 16×8 partitions or four 8×8 partitions. Each macroblock partition of size 8×8 could be further divided by single 8×8, two 8×4, two 4×8 or four 4×4 sub-macroblock partitions. Coding modes support variable block sizes that divide the frame into non-overlapping blocks. Every block requires signalling its coding information that includes reference frame index and motion vector[2]. The selection of block size is linked to complexity degree of the region. Macroblock partitions are used to predict areas with smooth texture variation while sub-macroblock partitions are used to provide accurate predictions for regions with high degree of variations. Skip mode is

[2] It could point to integer, half or quarter-sample (it will be discussed later in this subsection).

the most important coding mode that predicts the entire macroblock 16×16, where the encoder sends only a flag instead of transmitting prediction information and residual signal (Nukhet & Tunali, 2005; Marpe et al., 2006a).

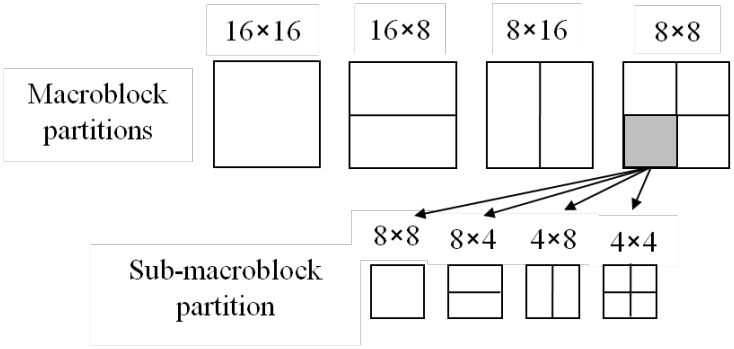


Figure 2-3 H.264/AVC coding modes (Ostermann et al., 2004)

H.264/AVC supports sub-pixel ME/MC[3]. The reference frame presents the integer sample (integer-pel) that might not be accurate enough during inter-picture prediction. Therefore, the codec generates half-pixel (half-pel) and quarter-pixel (quarter-pel) samples for each reference frame, where quarter-pel gets higher prediction accuracy than half-pel samples. Figure 2-4 depicts sub-pixel samples generation, where LPF is low pass filter. When a reference frame is stored in DPB, sub-pixel samples are generated, where integer samples are used to get half-samples that include horizontal and vertical samples. These samples are obtained through up-sampling; the reference frame then uses neighbouring horizontal or vertical integer samples (six samples) to interpolate half-pel sample through AVC Finite Impulse Response (FIR) filter. The filter has 6 tap with weights of (1/32, -5/32, 5/8, 5/8, -5/32, 1/32). Diagonal half-pel samples are obtained using either horizontal or vertical neighbouring half-pel samples. Quarter-pel samples are obtained through averaging the closest two samples; integer-pel and half-pel samples (Richardson, 2010).

Since each macroblock is handled separately, blocking distortion will be significant when coding the given frame at low bitrate. These types of artefacts appear as visible discontinuities among block boundaries. The codec deploys Deblocking filter (named loop filter) to minimise the effect of these artefacts by filtering the decoded block just before storing/displaying. Therefore, the blockiness artefacts are reduced which improves video quality perception and enhances inter-picture prediction when the filtered frame is used later as reference frame (Marpe et al., 2006a). It filters up to 3 pixels from each block side that would be in horizontal or

[3] Motion compensation constructs the predicted macroblock using prediction information.

vertical direction. The filter has four strength levels starting from strongest level that targets intra-predicted blocks to weakest level for inter-predicted blocks that have the same reference frames, motion vectors and no coded coefficients (Richardson, 2010).

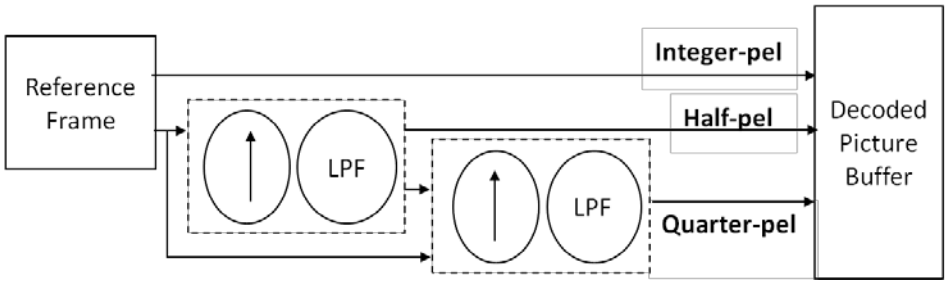


Figure 2-4 Sub-pixel samples generation via H.264/AVC

2.1.1.3 Block matching process

Block matching is the main process in inter-picture prediction. It searches for best block matching to the current macroblock, where the position of the predicted block that belongs to the reference frame is referred to as Motion Vector (MV). This is the actual motion vector that contains a pair of (x, y) coordinate that points to the predicted block itself. The codec predicts the actual MV through getting the median value among motion vectors that belong to neighbouring macroblocks (MV_p). The difference vector (Motion Vector Difference, *MVD*) is obtained by subtracting predicted motion vector from actual motion vector. Figure 2-5 shows block matching process using a number of coding modes among five reference frames. Macroblock partition 8×8 will be further divided into sub-macroblock partitions, where block matching will check these modes as well. Predicting P-frame using single reference frame requires 259 checks to cover all combinations for coding modes, where 256 checks is the combinations using sub-macroblock partitions (4^4) in addition to three checks for the macroblock partitions (16×16, 16×8 and, 8×16).

The motion estimation for a single and five reference frames would consume about 50% and 80% of the codec's total encoding time (Chiang et al., 2011; Xu & He, 2008). Therefore, motion estimation is the most computationally complex process of H.264/AVC. Searching for the best block matching requires computing visual distortion and amount of coded data. It includes motion vector difference, coding modes and reference frame index alongside the residual signal that is represented by transformed coefficients (it will be explained in the following subsections). This process is linked to rate-distortion optimization to guarantee the minimum cost for the final selection of block matching. It is a compromise between

the actual bitrate and distortion. The cost function is based on the Lagrangian method, *J(ref | λ_{Motion})*. It is defined by equation 2-1 (Jung et al., 2012),

$$J(ref|\lambda_{MOTION}) = SAD(s,r) + \lambda_{Motion} \times R(MVD, REF) \quad (2\text{-}1)$$

where Sum of Absolute Difference (*SAD*) frame is the prediction error between the current (*s*), and, corresponding reference block (*r*), λ_{Motion} is Lagrange multiplier, *R* is the number of bits required to code both; Motion Vector Difference (*MVD*) and reference frame (*REF*). The overhead cost of signalling the reference frame for each macroblock is related to the index position of the reference frame inside these buffers, where fewer number of bits are used to address the closer reference frames (e.g. the closest two reference frames requires single and three bits respectively). Therefore, H.264/AVC applies reference frames reordering when IPPP coding structure (it will be explained in the following subsection) is used to reduce the signalling of reference frames, where the closer reference frames have lower indexing value than further reference frames. This is accomplished through sorting decoded reference frames in descending order, where the index of nearest decoded reference frame (recent temporal frame) will be the first element in List 0 buffer (Shen et al., 2007).

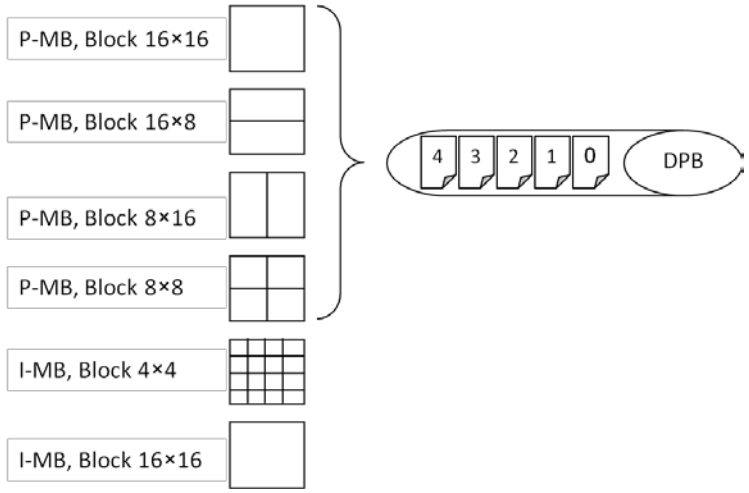


Figure 2-5 Block matching process when number of reference frames is 5 (Yu-wen et al., 2006)

2.1.1.4 Prediction architectures

Prediction architectures are defined through Reference Frame Selection (RFS) and Reference Frame Ordering (RFO). RFS identifies a set of reference frames, where they are stored in DPB. Reference frame ordering (RFO) defines how these frames' indices are placed inside the List buffer. Different combinations for RFS and RFO

lead to deriving different prediction architectures that eventually affect the coding performance of H.264/AVC.

There are two different coding structures (IPPP and IBBP); that are used among any prediction architectures. The former relies on I-frame and P-frame irrespective of frames' referencing selection. It has always low computational complexity and memory consumption with respect to IBBP as it allows forward prediction only. The latter deploys all frames' types in order to provide efficient coding gain at the expense of increasing both computational complexity and memory. This is due to deployment of block matching using forward, backward and bi-prediction in addition to storing both forward and backward reference frames (Richardson, 2010).

2.1.1.5 Statistical analysis of block matching

Reference frames have different roles of block matching depending on temporal correlation. Statistical analysis of block matching is a powerful technique to understand how much each reference frame contributes in inter-picture prediction.

Statistical analysis is conducted during block matching process using a set of counters that reflect the usage of each reference frames with different coding modes. The objective is to compute the amounts of the selected coding modes that includes the amounts of skip, 16×16, 16×8, 8×16 macroblock partitions alongside macroblock sub-partitions for each reference frame. When the block matching process determines the best coding mode for the current block, the corresponding counter for coding mode and reference frame is increased by one. When the given video is coded, the counters' values are processed in order to determine the amounts of inter-picture prediction for each reference frame. These values are first multiplied by a set of factors that reflect the size of each coding mode with respect to macroblock. E.g. the factor is one for skip and 16×16 block sizes, while it is two for 16×8 and 8×16 block sizes. The summation is then taken place for the amounts of coding modes that belong to the same reference frame. Normalisation is applied to define the amount of prediction for each reference frame in percent. E.g. the amounts of blocks using coding modes {16×16 and 16×8} among three reference frames are {820, 1150}, {550, 860} and {300, 590} respectively. The corresponding amounts with respect macroblock size would be {820, 575}, {550, 430} and {300, 295}. This entails that these reference frames have predicted 1395, 980 and 595 blocks, where the role for these frames are 46.97%, 33% and 20.03% respectively.

2.1.2 Transformation

The residual signal contains energy that has a high degree of redundancy that is exploited by the transformation. It de-correlates the energy to make it concentrate into a low number of coefficients. It concentrates the energy into few non-zero coefficients that are located around the DC coefficient while the opposite corresponding corner usually has zero coefficients (Richardson, 2010). There are two transforms that are based on Hadamard integer transformation; they are 4×4 and 8×8 integer transform. This type of transformation provides the codec two important properties. First it is reversible without any mismatch between encoder and decoder. Secondly, it provides easy hardware implementation through addition, subtraction and bit shifting (Sullivan et al., 2004; Richardson, 2010).

2.1.3 Quantisation

Quantisation is the main cause for visual quality degradation in a video codec, where transformed coefficients are scaled to a smaller set of values (Richardson, 2010). Therefore, it provides direct relationship among bitrate and video quality, where a high quantisation step provides significant coding ratio at the expense of significant distorted video quality. Quantisation exploits spectral redundancy, whereas the HVS is more sensitive to low frequency (colour intensity) than high frequency (edges). Therefore, the transformed coefficients located around the DC coefficient are quantised by a small factor while coefficients located at the opposite corner (represent edges) are quantised by a high factor. This non-uniform quantisation increases zero-coefficients while maintaining nearly the same visual perception (Richardson, 2010). Quantisation step size (QP) is a value among 52 values (scalar quantisation), where every 6 incremental step size reflects doubling quantisation (Sullivan et al., 2004). The transformed coefficients are scanned and placed into array after quantisation. The scan order starts with coefficients that are located around DC coefficient and continues toward high frequency coefficients (Richardson, 2010).

2.1.4 Entropy

Entropy encodes several elements that are generated through previous blocks. These elements are (Sullivan et al., 2004):

- Layer syntax including picture and slice header.
- Macroblock type involves prediction type and coding modes.
- Macroblock Coded Block Pattern (CBP) identifies which macroblock partition (8×8) contains non-zero transformed coefficients.

- Quantisation parameter sent as delta value from corresponding QP for previous macroblock.
- Reference frame indices that are used for inter-picture prediction.
- Motion vector that is signalled via *MVD*.
- Scaled transformed coefficients that correspond to residual signal.

The last element is the most dominant one when coding video at high bitrates while other elements are the main bulk of data when coding video at low bitrates (Sullivan et al., 2004). These elements have significant amount of statistical dependencies that are exploited by the entropy coder. It assigns short codes for frequent patterns and longer codes for irregular patterns. There are two entropy coders; they are Context Adaptive Variable Length Coding (CAVLC) and Context-based Adaptive Binary Arithmetic Coding (CABAC). The former is based on Huffman coding, where it codes only transformed coefficients. Twelve tables are deployed that describe coefficients number, magnitude and number of zero coefficients. The latter improves coding efficiency by 10% compared to CAVLC at the expense of more computational complexity (Sullivan et al., 2004). It has three main components, context modelling, binarisation and arithmetic coder. Context modelling selects the model according to observations from previously encoded blocks. The second component converts non-binary symbol into bins, which are coded by the last component (Richardson, 2010).

The codec has several coding tools that are usually not entirely used. Therefore, subsets of the supported coding tools are defined by the codec profile while its level identifies maximum limit of decoder capabilities (Richardson, 2010). Through configuring profile and level, H.264/AVC is used in a wide spectrum of applications. Detailed description for H.26/AVC are explained in these resources (Sullivan et al., 2004; Nukhet & Tunali, 2005; Marpe et al., 2006a; Richardson, 2010).

This section outlined key technologies behind H.264/AVC relevant to this thesis. These include multi-reference prediction, coding modes and sub-pixel ME and MC; that are deployed within the research investigations. The following section will introduce current extension for multi-view video codec (H.264/MVC) in terms of requirements, prediction architecture and its limitations.

2.2 H.264/MVC standard

H.264/MVC is introduced in this section, where its prediction architecture is highlighted. MVC was standardised for coding stereoscopic and multi-view video in 2008 (Chen et al., 2009b; Vetro et al., 2011; Dufaux et al., 2013). MVC is similar to

scalable video coding (SVC[4]) in structuring the videos into layers. It supports inter-layer prediction in order to exploit the inter-view dependency among these layers, where each layer represents a separate video.

2.2.1 Different scenarios for multi-view video coding

There are five main scenarios when displaying multi-view video at the receiver side, as reported in (Chen et al., 2009b), they are:

- Monoscopic display: shows single view. Therefore this extension is backward compatible to H.264/AVC's decoder.
- Stereoscopic display: presents two views. There is no head motion parallax and it is considered as the simplest form of multi-view video.
- Free-view point display: user selects a view among the received views.
- Narrow view angle display: supports few views to be displayed.
- Wide view range display: is capable of presenting a large number of views simultaneously.

2.2.2 Multi-view video coding general requirements

The following lists the generic requirements for MVC (Vetro et al., 2011):

- Coding efficiency: the codec should exploit the spatial redundancy among neighbouring views. It should outperform the coding efficiency of simulcast video coding; otherwise simulcast video coding[5] should be used to compress the given multi-view video.
- Backward compatibility: users who are provided legacy decoders, the coded multi-view should be compliant with them. Therefore the 1st layer (base layer) of codec should be decoded independently.
- Scalability: there are two types of scalability, they are namely:
  - Temporal scalability: displays the video through various frame rates.
  - View Scalability: multi-view displays have different capability in presenting views. For displays which support a limited number of views, view scalability is required. Therefore the encoder should not transmit extra views which will not be displayed at the decoder side.
- Random access: there are two types of random access, they are:
  - Temporal random access: the monoscopic video should be compressed in a way that supports decoding certain frames independently in order to preview video at different time slices.

[4] Its coded stream is decodable by users with different resources and network bandwidth
[5] Compress each view separately using monoscopic video coding

- View random access: in free-view television, a user chooses certain view, therefore the multi-view video should be carefully organised in a way to support minimum decoded frames between different views.

o Parallel processing: since multi-view video consists of a set of neighbouring views. Decoding the bitstream in a sequential manner would not be efficient solution for real-time applications. Therefore, the multi-view video codec should be designed in a way that supports parallel decoding to realise an acceptable decoding time for real-time applications (Ugur et al., 2007).

o Decoder's resources: the main critical resources are computational complexity and memory consumption. The multi-view video codec should be able to exploit the spatial redundancy among multi-view video without significantly increasing the decoder's resources because it might prevent displaying the video in a smooth way.

o Error resilient: error free transmission cannot be guaranteed when packet data might be lost. Therefore, multi-view video codec should enable robust transmission to the decoder especially for environments with an error-prone network.

2.2.3 Typical prediction architecture

Multi-view video coding exploits spatial redundancy among neighbouring views to improve coding efficiency with respect to simulcast video coding. The inter-picture mechanism for compressing monoscopic video has been extended in multi-view video codec. Hierarchical B-picture (HBP) is the most efficient prediction structure and it has evolved and become the typical prediction architecture for multi-view video coding (Zhang et al., 2011a). It is depicted in Figure 2-6, where V_i and T_i reflect the *view-id* and *temporal-id*; time slice numbers respectively (Chen et al., 2009b; Vetro et al., 2011). Base view is V_0, where it uses only temporal reference frames for inter-picture prediction, while the remaining views (odd and even views) are dependent views that use temporal and spatial reference frames. Odd views are V_1, V_3, V_5 and V_7 while V_2, V_4, V_6 are even views. Each group of pictures in base view has a key frame that could be either I-frame or P-frame. Odd views allow bi-prediction from time and view directions, e.g. B-frame that belongs to view V_1 (located at time slice T_4) is predicted using temporal frames (B-frames located at time slices T_0 and T_8) and spatial frames (B-frames belong to V_0 and V_2 at time slice T_4). There are a set of temporal levels within HBP prediction architecture, where level 1 has frames that located at time slices T_0 and T_8, while level 2 has frames located at T_4. Frames located at T_2 and T_6 are belong to temporal level 3, while frames located at T_1, T_3, T_5

and T_7 are belong to temporal level 4. This arrangement allows predicting B-frame that is located at a certain temporal level by frames located at a lower temporal level.

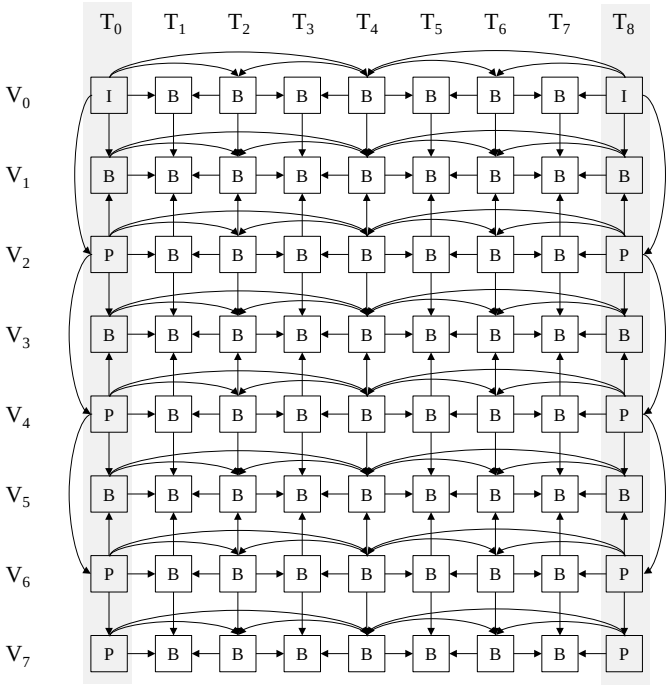


Figure 2-6 Typical prediction architecture for multi-view video coding (Jeon et al., 2009)

Multi-view video coding extension of H.264/AVC deploys symmetric coding (parameter settings are similar among neighbouring views; e.g. spatial-, temporal-resolution and quality). It's prediction architecture fulfils the requirements for generic multi-view video codec (Chen et al., 2009b).

- Backward compatibility: it is preserved through providing a base layer without any requirements for any reference frames from neighbouring views. At the decoder, the frames that belong to the base view are exploited without the need to decode frames that belong to neighbouring views.
- Scalability: temporal scalability is maintained through using hierarchical B-picture for each view. View scalability is achieved through determining a priority identifier. It reflects the views which will be transmitted. Therefore, it uses *view-id* in addition to *temporal-id*. Figure 2-7 shows four different operation points[6] in terms of *priority-id*, where the four different shaded grey levels correspond to these operating points. First operating point will display key frames (I_0, I_8, $I_{16,}$ etc.) that belong to V_0 at the lowest frame rate; 7.5 FPS while the highest operating point will display all frames that belong to V_0, V_1 and V_2 at 30 FPS.
- Random access: Instantaneous Decoder Refresh (IDR) clears the contents of DPB. It relies on intra-prediction which provides temporal random access through

[6] It is the combinations of temporal-id and view-id that will facilitate parsing coded bitstream.

dividing the video into a Group Of Pictures (GOP). Each GOP starts with I-frame (key or anchor[7] frame). When random access is requested, the decoder searches for the closest key frame to start decoding it independently followed by decoding next frames.

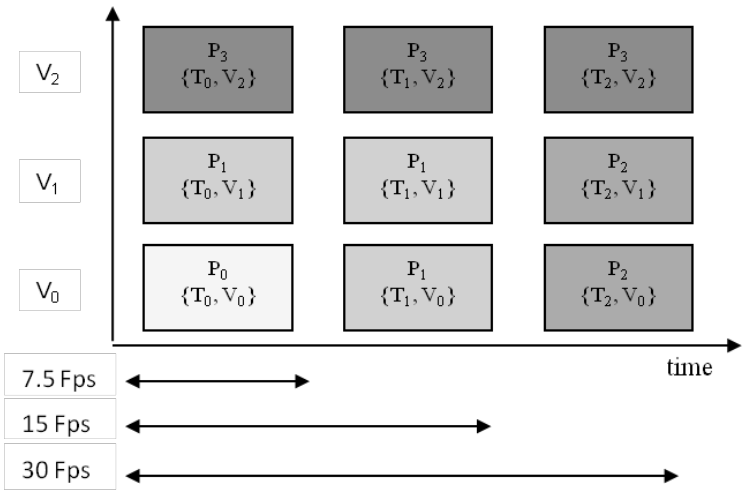


Figure 2-7 Different priority-id for three views (Chen et al., 2009b)

Frames need to be multiplexed into single stream prior to coding via multi-view video coding. There are two coding orders; view-first and time-first coding orders. The former multiplexes the frames that belong to certain view (they are located in the same GOP) then the frames that belong to the neighbouring view are then inserted. Time-first coding order multiplexes the frames that belong to neighbouring views in a sequential manner (belonging to the same time slice) then frames that belong to the next time slice are inserted afterward. Figure 2-8 illustrates time-first coding order that is a common order as it provides low decoding delay among views with respect to view-first coding (Chen et al., 2009b; Vetro et al., 2011). H.264/MVC operates from low to high bitrates depending on the application. Based on common test conditions which was defined through Call for Proposals (CfP), there are two types of coding conditions (Su et al., 2006). The first condition is coding multi-view sequence at constant quality; hence the common test conditions provide a set of quantisation parameters, where the video codec should follow it. The other condition is coding multi-view video at constant bitrate, hence there are three defined ranges; each reflects the amount of average bitrate per one view which is represented by the average Kilobits per second (Kbps). This range defines low, medium and high bitrates constraint for each multi-view video sequence.

[7] Anchor pictures also referred to frames that follow key frame at same time slice.

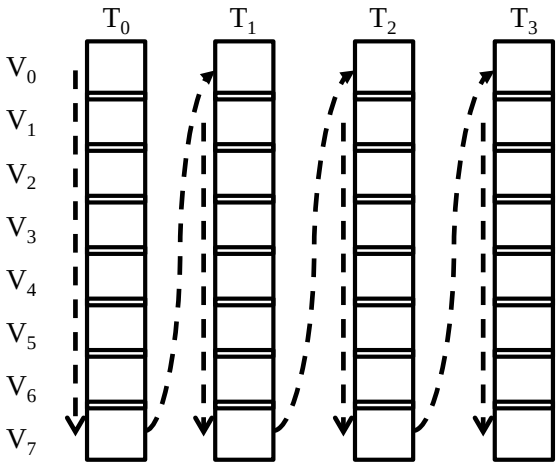


Figure 2-8 Time-first coding order (Chen et al., 2009b)

Block matching is applied similarly to the corresponding one in monoscopic video coding. The majority of frames in the prediction architecture are B-frames that imply using two buffers to store reference frames indices through List 0 and List 1 buffers. Predicted block is chosen based on rate distortion that considers all available reference frames and coding modes. The process of checking block matching using spatial reference frame that belongs to neighbouring view is disparity estimation, while the corresponding process using temporal reference frame is motion estimation. These predictions are named inter-view (disparity) and temporal prediction respectively. In the context of block reconstruction, the process of compensating it by temporal reference frame via motion vector is named motion compensation, whilst the corresponding process that uses spatial reference frame via disparity vector is disparity compensation. The best block matching has minimum cost that represents current block through the information; motion-vectors or disparity-vectors, coding mode and residual transformed coefficients.

2.2.4 Multi-view video coding limitations

The multi-view video coding extension, H.264/MVC, uses B-frames to achieve efficient coding performance at the expense of large computational complexity and memory requirements[8]. It puts restriction in the prediction architecture that includes RFS and RFO. Figure 2-9 shows successive frames that belong to four cameras, where each rectangle represents a frame. F, S and T represent current frame, spatial frame and temporal frame respectively. It enforces inter-view prediction using only

[8] This is due to enable backward and forward reference frames that increase amount of time for deploying block matching in addition to increase DBP size to store these reference frames.

frames that belong to the closest neighbouring views (S frames), where these frames are contained in the same access unit (Chen et al., 2009b; Vetro et al., 2011). Therefore, it does not support prediction using frames in spatiotemporal direction that belong to different access units. The multi-view video coding standard controls the RFO via parameter named *InterPredPicsFirst*. This parameter can select either placing spatial or temporal reference frames indices first in the List. Since the majority of the prediction comes across temporal direction, the codec uses temporal-first as a default setting (ISO/IEC MPEG & ITU-T VCEG, 2008). The challenge is to determine which source of reference frames (temporal or spatial) is more significant. It is difficult to decide the most significant source for frames that belong to odd views (located at low temporal levels) especially when GOP size is large, e.g. B-frame located at time slice T_8 when GOP size is 15.

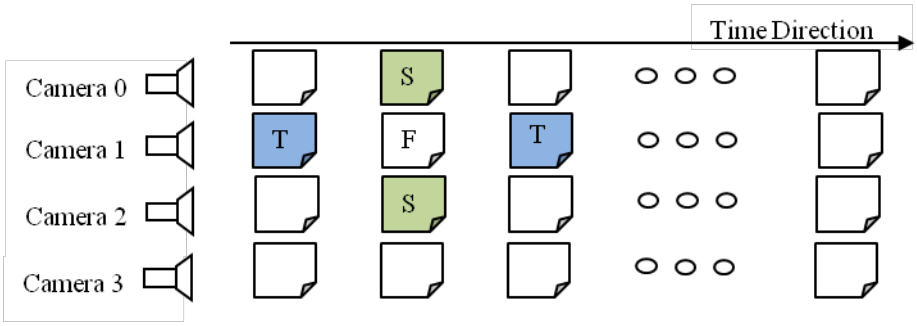


Figure 2-9 Temporal and spatial prediction via H.264/MVC

The prediction architecture is what distinguishes multi-view video coding from simulcast video coding through allowing inter-picture prediction from different directions. On the other hand, prediction information is the main bulk of data when the codec operates at low bitrates. This entails the significant importance of prediction architecture when the codec operates at low bitrates. H.264/AVC is the base codec for multi-view video coding extension that supports IPPP and IBBP coding structures. It provides more flexibility for deriving prediction architectures in terms of reference frame selection and reference frame ordering. Therefore, H.264/AVC is used during the research investigations.

This section introduced multi-view video coding extension, where different display scenarios, requirements and typical prediction architecture are illustrated. The justification behind selecting H.264/AVC instead of H.264/MVC is clarified. The next section will outline video quality metrics, where the objective quality metrics that are used in this research are highlighted.

2.3 Video quality metrics

A Video Quality Metric (*VQM*) measures the amount of quality degradation in coded video. There are several types of coding artefacts that exist in coded video. The most common artefacts are blockiness and blurriness (Boev et al., 2011b). Blockiness appears when coding video at low bitrate due to scaling the transformed coefficients coarsely, while blurriness results from interpolating frames, where high frequency components are degraded.

Video quality metrics are categorised into subjective and objective video quality metrics. Subjective metrics rely on assessing the visual quality through a group of viewers who judge the quality through watching reference video (un-coded) and coded videos (impaired). Although video quality measures obtained by subjective metrics are more reliable than objective video quality metrics, they are costly and need more time to be conducted. Significant time is needed for setting up the laboratory in a controlled lighting condition and performing tests among all assessors prior to conducting subjective assessment, e.g. visual acuity test, colour vision test and stereo vision test (Boev et al., 2011b; Pedro & Velasco, 2012).

Objective video quality metrics use predefined formulas to determine the quality of coded video. There are three types of objective metrics depending on availability of reference video; they are: full reference, reduced reference and no reference metrics (Richardson, 2010; Pedro & Velasco, 2012). Full reference metrics require full availability of reference video, while reduced reference metrics use certain characteristics of reference video which are sent as side information beside coded video. No reference metrics assess the quality degradation of coded video without the need of reference video.

Pedro and Velasco categorise full reference metrics according to the methodology deployed for each metric (Pedro & Velasco, 2012); they are:

- Pixel-based metrics: the most time efficient methods that measure the degradation at pixel level, such as Mean Square Error (*MSE*), Signal-to-Noise Ratio (*SNR*) and Peak Single-to-Noise Ratio (*PSNR*) metrics. *PSNR* metric is the most common used in video coding due to its simplicity to compute video quality. It is combined with bitrate to measure Rate-Distortion curve (R-D) for the coded video.

$$PSNR = 10\, log_{10} \frac{(2^D - 1)^2}{MSE} \tag{2-2}$$

$$MSE = \frac{1}{M.N} \sum_{i=0,\ j=0}^{M-1,\ N-1} (X_{i,j} - Y_{i,j})^2 \quad (2\text{-}3)$$

where *MSE* is mean square error while *X* and *Y* are luminance components of un-coded and impairment frame respectively. *M*, *N* and *D* are horizontal, vertical dimensions and pixel bit depths (Richardson, 2010; Pedro & Velasco, 2012).

In context of mixed spatial-resolution frames, the coded low spatial-resolution frames are interpolated prior to measuring *PSNR*. This metric has two measures; they are actual and over-estimated *PSNR* measures. Figure 2-10 shows these measures, where ENC, DEC and LPF are encoder, decoder and low pass filter respectively. F_{org}, F_{org}' and F_{coded} refer to un-coded frame (ground truth), un-coded frame (interpolated) and coded frame (interpolated) respectively. Computing *PSNR* from F_{org} and F_{coded} is $PSNR_{actual}$, that measures coding and blurriness distortions. $PSNR_{over\text{-}estimated}$ is computed using F_{org}' and F_{coded}. It measures amount of coding distortions only. Therefore its measurement is higher than $PSNR_{actual}$. Since majority of studies rely on $PSNR_{over\text{-}estimated}$, it is used as an objective quality measurement in chapter five.

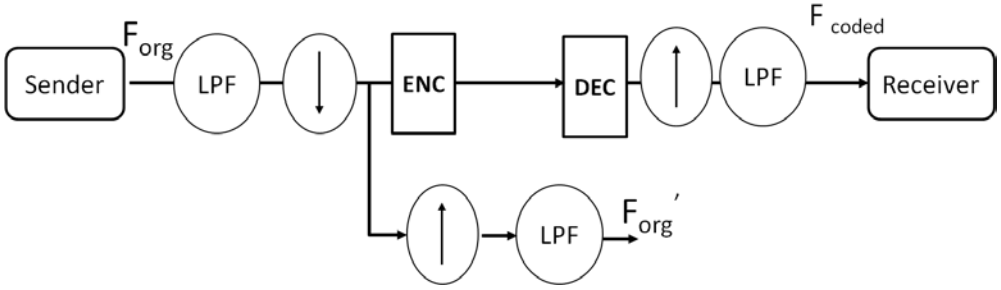


Figure 2-10 Actual and over-estimated Peak Signal-to-Noise Ratio

o Based on structural similarities: they evaluate quality degradation for coded videos through measuring different aspects of images that affect HVS. One of the popular methods is Structural SIMilarity index (*SSIM*). It measures the perceived changes through the image's structural information that satisfies three conditions. They are symmetry (*SSIM(x, y) = SSIM(y, x)*), boundedness (*SSIM(x, y) ≤1)*) and unique maximum (*SSIM(x, y) = 1* if and only if *x = y*). It performs luminance, contrast and structure comparisons, where *SSIM* is the linear combination of these measurements (Wang et al., 2004).

$$MSSIM(X,Y) = \frac{1}{W}\sum_{j=1}^{W} SSIM(x_j, y_j) \quad (2\text{-}4)$$

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (2\text{-}5)$$

where *MSSIM* is average *SSIM* that represents image quality. *W,* μ_x, μ_y are the number of local windows in image, mean intensity for pixels in horizontal and vertical directions respectively. σ_x and σ_y are standard deviation for pixels in horizontal and vertical directions while C_1 and C_2 are two constants to avoid instability when denominator is zero (C_1 and C_2 are set to 6.5 and 58.5 respectively) (Wang et al., 2004).

- Based on artefacts: these types of metrics aim to measure amount of different artefacts such as blockiness, blurring and ringing. One of the most popular metric is *Lee et al.* metric which has been recommended via ITU-T (ITU-T, 2004, 2008; Lee et al., 2011). The metric measures the video quality degradation in areas around edges which affect significantly perception of HVS. The metric measures edge Peak Single-to-Noise Ratio (*EPSNR*), blockiness and blurriness, where their linear combination represent video quality measurement for this metric. First, the edges are extracted from the reference image using horizontal and vertical gradient operators followed by threshold to obtain the mask for a given image. This mask defines the pixels among the reference and impairment images that will be used to calculate *EPSNR*. Reference and impairment images are high pass filtered via SOBEL gradient operator to obtain their horizontal and vertical gradient images. These images are used to compute blockiness and blurriness. The next equations depict how to compute this Video Quality Metric (*VQM*).

$$VQM = EPSNR + w_1 \times F_{blocking} + w_2 \times F_{blur} \quad (2\text{-}6)$$

$$EPSNR = 10\, log_{10}\left(\frac{P^2}{MSE_{edge}}\right) \quad (2\text{-}7)$$

$$F_{blocking} = \frac{1}{n_{blocking}}\sum_{k} HV_p(k) - HV_s(k)\ if\ \left(HV_p(k) > HV_s(k)\right) \quad (2\text{-}8)$$

$$F_{blur} = \frac{1}{n_{blur}} \sum_{k} HV_s(k) - HV_p(k) \; if \; \left(HV_s(k) > HV_p(k)\right) \tag{2-9}$$

$$HV(t,i,j) = \begin{cases} R(t,i,j), & \begin{matrix} R(t,i,j) \geq r_{min} \quad and \\ m\frac{\pi}{2} - \Delta\theta < \theta(i,j,t) < m\frac{\pi}{2} + \Delta\theta \end{matrix} \\ 0, & otherwise \end{cases} \tag{2-10}$$

$$R(t,i,j) = \sqrt{H(t,i,j)^2 + V(t,i,j)^2} \tag{2-11}$$

$$\theta(t,i,j) = tan^{-1}\left[\frac{V(t,i,j)}{H(t,i,j)}\right] \tag{2-12}$$

where $F_{blocking}$ and F_{blur} are the amount of blockiness and blurriness. The constant numbers in equation 2-6 (w_1 and w_2) are set to -1/14 (Lee et al., 2011). In equation 2-7, P and MSE_{edge} represent maximum value in image (255 for pixel depth of 8-bits) and mean square error for pixels identified through the mask (gradient image after thresholding). Horizontal and vertical component for reference and impairment images are HV_s and HV_p respectively. $n_{blocking}$ and n_{blur} are number of pixels that satisfy the conditions in equations 2-8 and 2-9 respectively. In equation 2-10, r_{min} and $\Delta\theta$ are two thresholds that are set to 110 and 0.225 respectively (Lee et al., 2011). *H, V, R* and θ are horizontal, vertical gradient images, gradient magnitude and gradient direction (angle) respectively.

- Based on the vision model used: these types of metrics simulate certain models that replicate the stimulus of HVS perception in order to estimate the video quality close to actual perception. These models use datasets in training phase in order to define model parameters values. *De Silva et al.* proposed Stereoscopic Structural Distortion (*StSD*) that uses suppression theory to quantify the visual quality for mixed spatial-resolution stereoscopic videos (De Silva et al., 2013). This metric measures both structural distortion and asymmetric blur. The former measures amount of changing objects' structures while the latter is quantified by the amount of image sharpness degradation. Structural distortion is measured by decimating the frame by a factor of two in horizontal and vertical directions. Frames from right and left views are partitioned into a number of blocks, where block size is 13×13 pixels. Structural difference is then computed using reference and impairments frames. The resulted structural distortions (d_s) from right and left views (d_R and d_L respectively) are summed to get final structural distortion. On the

other hand, Blurriness artefact is computed based on the magnitude difference for the reference and impairment frames' edges that are previously extracted by SOBEL filter. The asymmetric blur (*B*) is the minimum blur among right and left views (b_R and b_L respectively) according to suppression theory (De Silva et al., 2013). The following equations are used to compute *StSD* metric.

$$StSD = \frac{0.7343}{1 + e^{(-15.778.(d_s - 0.14))} - 0.073 + 0.0085.B} \tag{2-13}$$

$$d_s = d_L + d_R \tag{2-14}$$

$$d = 1 - (0.5.d_m + 1.5.d_h) \tag{2-15}$$

$$B = min(b_L, b_R) \tag{2-16}$$

$$B_{i,j} = \begin{cases} \Delta e_{i,j} & \Delta e_{i,j} > \frac{\sigma_{S_o}}{2} \\ 0 & \Delta e_{i,j} \leq \frac{\sigma_{S_o}}{2} \end{cases} \tag{2-17}$$

$$\Delta e_{i,j} = \begin{cases} S_o(i,j) - S_c(i,j) & S_o(i,j) > \overline{S_0} \\ 0 & S_o(i,j) \leq \overline{S_0} \end{cases} \tag{2-18}$$

where d_m and d_h are the mean and highest structural distortion, while σ_{So} and $\Delta e_{i,j}$ are standard deviation for edge magnitude of un-coded frame and edge magnitude difference for reference and coded frames. $\overline{So}$, S_0 and S_c *are* the average of edge magnitude, edge magnitude for reference and coded frames respectively.

2.4 Chapter Summary

This chapter described briefly H.264/AVC, where the prediction component is highlighted. Parts of the coding tools that are supported by prediction are multi-reference prediction, coding modes, sub-pixel ME and MC. These coding tools are studied during the research investigations. Multi-reference prediction supports inter-picture prediction from multiple frames. Variable block size provides multiple block partitions that suit blocks with different complexity (homogenous and detail contents).

Sub-pixel ME and MC provide higher degree of accuracy than integer-pixel samples by interpolating the reference frame at a level of half and quarter-pixel. The chapter presented a multi-view video coding extension (H.264/MVC) that provides backward compatibility with legacy decoders, view scalability and random access at the expense of high computational complexity and memory requirements. This extension puts constraints on reference frame selection and reference frame ordering. Since H.264/AVC supports greater flexibility on both RFS and RFO, it is used during the investigations reported in this thesis. The Chapter also outlined video quality metrics, and highlighted objective video quality metrics that are used during the research investigations.

The next chapter provides taxonomy for low bitrate video codecs. Symmetric multi-view video coding and asymmetric spatial-resolution multi-view video coding are then reviewed, where the challenges that are addressed in this thesis are summarised.

CHAPTER 3. REVIEW OF H.264 BASED MULTI-VIEW VIDEO CODING

This chapter reviews coding approaches that are used for multi-view video coding based on H.264 at low bitrates. It includes resolution-based, depth-based, model-based and hybrid-based coding approaches. Prediction architectures and block matching efficiency are then discussed for symmetric MVC. After that, prediction architectures and visual enhancement algorithms are reviewed for mixed spatial-resolution MVC. A summary of the review is then outlined, followed by a summary of the investigations that are addressed in this thesis.

3.1 Low bitrate video codecs

This section reviews different coding approaches that are conducted when coding multi-view video at low bitrates. The coding approaches are categorised based on the key criteria that identifies each coding solution. Part of the coding approaches relies on resolution-based, where they either use similar setting or different coding parameters settings. Another category relies on integrating depth-maps with texture views in order to compress subset of views. Multi-view video could be coded using model-based, where either object or mesh-based coding is applied. The last approach integrates different combinations (hybrid-based) from the previous coding approaches. Figure 3-1 presents the taxonomy for coding multi-view video at low bitrates, where these coding approaches will be briefly outlined in the following subsections.

3.1.1 Resolution-based approach

The resolution-based coding approach is classified into either symmetric or asymmetric. The symmetric coding approach uses similar settings for all views that include spatial-resolution, temporal-resolution and quality, while asymmetric-based coding deploys different coding parameters settings among neighbouring views.

3.1.1.1 Symmetric Coding

The quantisation parameter is a straightforward solution to reduce bitrate, where the codec increases the quantisation parameter in order to meet the target bitrate. This entails transforming the residual coefficients coarsely. Symmetric coding is preferred when the video quality for the dependent view is in range of 28 dB and 32 dB while mixed spatial-resolution is preferred when the corresponding video quality for the

dependent view is below 28 dB (Saygili et al., 2010). When dependent view is coded below the threshold (32 dB), symmetric coding obtains better quality than asymmetric quality coding approach for stereoscopic video coding (Saygili et al., 2011). Symmetric coding and video plus depth are preferable coding solutions than simulcast video coding and asymmetric spatial-resolution[9] for mobile 3D television (Tech et al., 2009b; Strohmeier & Tech, 2010). Blocking artefacts is the main challenge for this coding approach, where it appears as large discontinuity distortions among nearby blocks. Truncating a set of views and coding remaining views with their depth maps provides better coding solution than symmetric coding at low bitrates, where accurate depth maps are crucial for the quality of synthesised views (Savas et al., 2011, 2012). Symmetric coding is one of the potential solutions for low bitrates, where the codec does not need additional amendments compared to other coding approaches. It requires neither depth map nor modelling the scene as in both depth-based and model-based coding approaches.

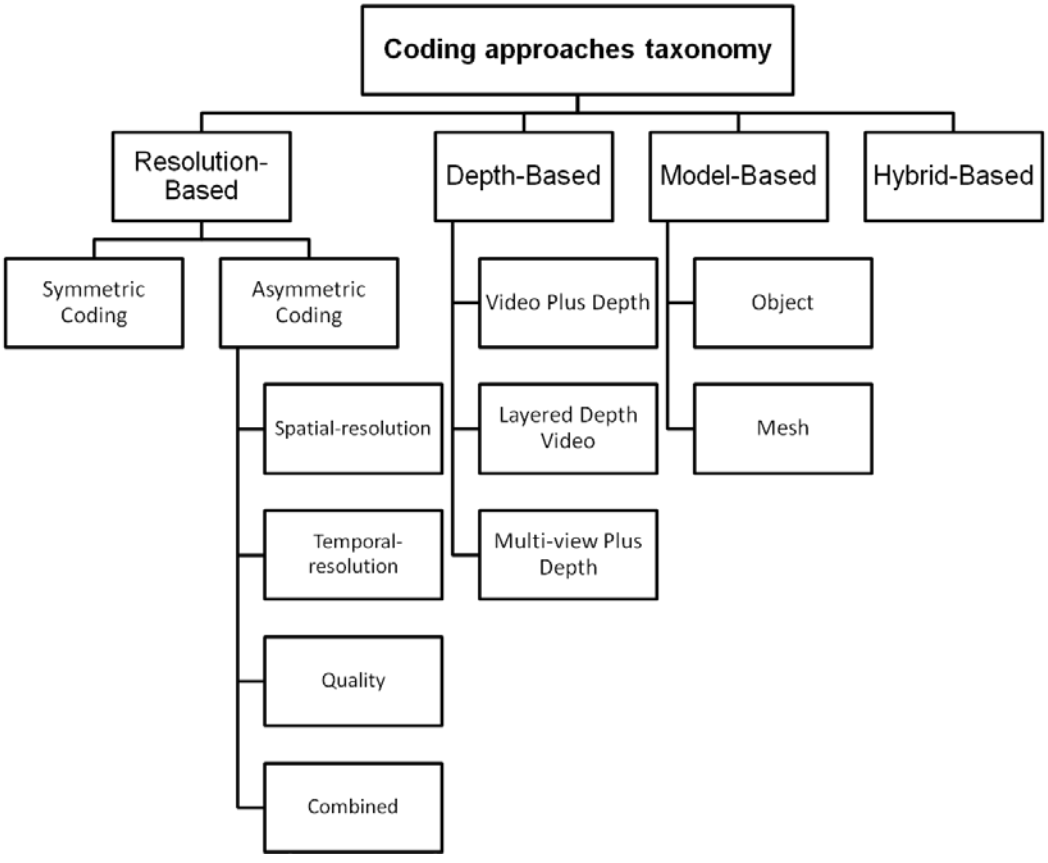


Figure 3-1 Coding approaches for low bitrate applications

[9] Asymmetric spatial-resolution has been applied using simulcast video coding

The efficiency of symmetric coding approach is linked to the amount of spatial correlations among neighbouring views. H.264/AVC exploits these redundancies to provide superior coding gain than simulcast video coding at low bitrates (Merkle et al., 2007a). Since multi-view video coding requires higher computational complexity than simulcast video coding, it should not be used when the amount of spatial redundancies among neighbouring views are insignificant. Although several studies looked into block matching efficiency, they do not define clear criterion for the best usage of multi-view video coding (Merkle et al., 2007a; Bouyagoub et al., 2010). Prediction architecture is the core part of multi-view video coding. Although several prediction architectures have been proposed for symmetric coding, they provide neither sufficient justification behind their prediction architectures nor propose practical solution that fits low bitrate applications (Zhang et al., 2008; Pourazad et al., 2009a; Seungwook & Yang, 2011).

3.1.1.2 Asymmetric Coding

Asymmetric coding is the second type for resolution-based coding approach. It reduces the amount of data prior to compression in order to provide potential coding solution to symmetric coding approach.

3.1.1.2.1 Asymmetric spatial-resolution

Asymmetric spatial-resolution video coding relies on resolution reduction which entails coding fewer amounts of visual data with respect to symmetric coding. This leads to significantly improving coding efficiency at low bitrate. This approach has been used in monoscopic video coding, where lower spatial-resolution of the input frames are coded and transmitted. In the decoder side, higher spatial-resolution frames are generated using frame enlargement techniques (Uslubas et al., 2010; Tech & Babu, 2011). This concept has been extended to stereoscopic video coding, where frames spatial-resolution, that belong to certain view are reduced, while frames that belong to other view are maintained in their full spatial-resolution. This technique is called mixed spatial-resolution stereoscopic video coding which relies on suppression theory. This theory states that the perceptual quality of stereoscopic video would be close to the higher quality view where 30% to 35% of bitrate is advocated to the view that has low spatial-resolution frames (Brust et al., 2009). Asymmetric spatial-resolution compromises blocking with blurring artefacts in order to achieve better rate distortion than symmetric video coding (Fehn et al., 2007). The perceived quality for mixed spatial-resolution stereoscopic video coding has been reported to be close to quality of the higher spatial-resolution view (Aflaki et al., 2010,

2013a). At low bitrates, mixed spatial-resolution coding approach reduces blocking artefacts in addition to reducing encoding and decoding complexities compared to symmetric coding (Brust et al., 2009; Aflaki et al., 2013a).

Few studies looked into investigating prediction architecture for this coding approach. They either inherited prediction architecture from symmetric coding or proposed prediction architectures in the context of stereoscopic video coding (Chen et al., 2008a; Fehn et al., 2007). Mixed spatial-resolution stereoscopic video by simulcast video coding is found to provide inferior quality compared to symmetric and video plus depth. On the other hand, mixed spatial-resolution could be applied in free-viewpoint TV (FTV). Therefore, few studies addressed enhancing visual quality for interpolated frames where they do not offer an efficient visual quality solution that fits low bitrate applications (Tech et al., 2009a; Najafi, 2012).

3.1.1.2.2 Asymmetric temporal-resolution

Asymmetric temporal-resolution reduces number of frames prior to coding through dropping these frames (temporal filtering), after the decoding process; the skipped frames are interpolated through the neighbouring decoded frames (Aksay et al., 2006). It can be efficiently implemented by a lifting scheme (pyramid decomposition) (Ozbek & Murat Tekalp, 2006). The input sequence is split into two streams, where one sequence is used to predict the other (prediction phase). After that, the residual signal is obtained via subtracting the second stream from the predicted one and an update is applied to increase the smoothness of the next prediction step. This approach is implemented through Motion-Compensation Temporal Filtering (MCTF) for monoscopic video coding (Schwarz et al., 2006). In multi-view video coding, there is view dimension, therefore, there are two decomposition directions that are conducted through MCTF and Disparity-Compensated View Filtering (DCVF) (Yang et al., 2006; Garbas et al., 2011).

The asymmetric temporal-resolution approach causes flickering artefacts (jerky appearance in terms of sharpness and quality) for sequences that contain fast objects' motion (Stelmach et al., 2000; Yea & Vetro, 2009). This negatively affects the visual perception for this coding approach at the receiver side.

3.1.1.2.3 Asymmetric quality

Asymmetric quality applies different quantisation step sizes among neighbouring views (Shafique et al., 2010). Different quality in stereoscopic video coding yields to average quality perception when subjectively assessed (Palaniappan & Nikil, 2012; Aflaki et al., 2013a). This concept has been extended to multi-view video, where

even views are coded in high quality while odd views are compressed in low quality in order to provide different qualities for the neighbouring views (Shafique et al., 2010). This approach allows great flexibility to achieve bitrate adaptation without demanding major amendments on the video codec (Gurler & Tekalp, 2013). This is achievable through applying different levels of quantisation parameters to reach target bitrate. It has been reported that degrading lower quality view by less than this threshold (31 dB and 33 dB for parallax barrier and projection displays respectively) increase blockiness artefacts, where perceived quality will be closer to lower quality view (Saygili et al., 2011).

Generally, asymmetric quality provides great flexibility among other asymmetric approaches. The blockiness artefacts would be an inevitable obstacle when coding multi-view videos at low bitrates (Savas et al., 2012). Symmetric coding and asymmetric spatial-resolution provide better perceived quality than asymmetric quality at low bitrates (Saygili et al., 2010; Aflaki et al., 2013a).

3.1.1.2.4 Combined asymmetric

The combined asymmetric coding approach integrates pervious asymmetric coding approaches for stereoscopic video coding. Several combinations have been generated using asymmetric settings for spatial-resolution, temporal-resolution and quality. Seven and six combinations have been proposed in *Ozbek et al.* and *Eichhorn & Ni* investigations (Ozbek & Tekalp, 2008; Eichhorn & Ni, 2009). The first study confirmed the acceptable visual quality when asymmetric spatial-resolution and quality are deployed while the second study concluded that the visual degradation using different asymmetric coding is more dependent on sequence, e.g. low motion sequences prefer asymmetric quality and asymmetric spatial-resolution coding. Asymmetric spatial-resolution is combined with asymmetric quality such that low spatial-resolution frames use lower QP than full spatial-resolution frames (Brust et al., 2010; Aflaki et al., 2013a).

3.1.2 Depth-based approach

In this category, the video is attached with its corresponding depth map, where it reflects the objects' distance from the camera. This additional information is used when novel views are constructed. Through depth maps, subsets of texture-views are sent instead of transmitting all texture videos. The size of depth map (grey-scale) is much less than texture video[10], therefore the total size of raw data decreases

[10] Depth's optimal Bitrate is 10 % to 30% for Video Plus Depth coding approach (Tech et al., 2009b)

significantly when compared to multi-view video with texture coding format. Video plus depth, layered depth video and multi-view plus depth provide a trade-off between synthesised view's quality and total bitrate.

3.1.2.1 Video plus depth

Video plus depth is the simplest subcategory in depth-based approach. In this coding approach, single texture video is accompanied with its depth map. This coding approach has proved its superior coding efficiency when an assessment is carried out using H.264/AVC which compresses texture and depth map separately (Merkle et al., 2009b). It saves up to 50% and entails higher coding efficiency than simulcast video coding at low bitrate, where acceptable quality level is dependent on depth map quality and scene content. For a smooth depth map with low structure complexity, Video plus depth has shown its superior subjective score when it is compared to simulcast video coding and mixed spatial-resolution stereoscopic video coding (Tech et al., 2009b; Strohmeier & Tech, 2010). The quality of rendered view is affected by the quality of the estimated depth map and the amount of disocclusion[11], where inaccurate depth map and large amounts of disocclusion cause annoying artefacts for the rendered view at the receiver side (Kauff et al., 2007; Oh et al., 2009).

3.1.2.2 Layered depth video

Layered Depth Video (LDV) tackles disocclusion that arises in video plus depth through using extra information (occlusion layers). In this coding approach, single texture is associated with depth map in addition to occlusion texture and occlusion depth layers. The last two layers are used during rendering; hole filling (Tian et al., 2009). The occlusion (residual) layers are obtained by warping the texture layers and subtracting it to determine the occluded areas (Barsi et al., 2008). This coding approach has higher computational complexity than Video Plus Depth since the residual layers are generated at the sender side (Daribo & Saito, 2011).

[11] It reflects areas that does not exist in the reference frame while occlusion reflects areas that exist only in the reference frame (Oh et al., 2009)

3.1.2.3 Multi-view plus depth

Multi-view plus depth (MVD) relies on attaching more texture views with their depth maps in order to facilitate better rendering for intermediate views. This coding approach does not use residual layer (occlusion layers) as in the layered depth video. Depth maps require on average 40% to 60% of total bitrate (Bosc et al., 2011). Multi-view plus depth usually select the outmost views; left and right in addition to central view with their corresponding depth maps in order to support free view in wide range navigation (Jeon et al., 2009; Liu et al., 2011; Savas et al., 2012). Therefore, MVD is more suitable to autostereoscopic displays than video plus depth and layered depth video. Both texture and depth maps are coded together such that each texture frame is followed by its depth map in the same access unit (Hannuksela et al., 2013).

3.1.3 Model-based coding approach

Model-based coding covers object-based and mesh-based coding approaches. They are both content-based coding, where an analysis is required prior to compression.

3.1.3.1 Object-based coding

The object-based coding approach processes the frame prior to compression through extracting the objects from the background. The background image is coded separately alongside extracted foreground objects, where the binary mask[12] is sent as side information. This is different from the conventional compression (hybrid video coding) which divides the frame into blocks, where each block is processed separately from its neighbours. At low bitrate, object-based coding approach does not suffer from blocking artefacts (Belloulata & Zhu, 2007). Background objects that exist in successive frames are represented by sprite (large background image obtained through camera motion parameters). Foreground objects are segmented by binary mask. At the decoder side, the binary mask is used to composite foreground objects with sprite in order to reconstruct views (Krutz et al., 2007; Wei, 2007). Segmentation is applied for intra-frame and inter-frame in order to provide efficient solution for object segmentation using H.264/AVC (Narasak et al., 2008). Background sprite could be single or multiple. Single sprite combines background objects that exist in successive frames into a single frame whilst multiple sprite generates set of partitions for these objects. This improves coding efficiency especially for sequences that are captured by cameras with large pans (Krutz, 2010). Object-based coding is generally efficient solution when coding sequences that

[12] It is used to enable foreground objects composition with sprite sequence (Krutz et al., 2007).

contain background objects more than foreground objects, where the encoder transmits the coded sprite once in addition to transmitting the coded foreground objects for every frame.

3.1.3.2 Mesh-based coding

The mesh-based coding approach offers great support for scene navigation as in free viewpoint video. Objects are represented using a mesh model with corresponding texture. In each view, objects are segmented, where 3D objects are presented as voxel model. Texture is extracted from an object's surface and presented as 3D mesh. Dynamic mesh is deployed in multi-view video, where triangles have connectivity over time (Smolic et al., 2007). In reconstruction, voxels are projected while texture is weighted from closer cameras (Smolić & Kauff, 2005; Smolic et al., 2006). It needs accurate segmentation for the object of interest to be reconstructed in high quality (Smolić & Kauff, 2005). H.264/AVC is used to encode texture information, while a 3D model is coded and sent to the receiver, where optimal source for each patch is selected and mapped to 3D object model (Chiang et al., 2012). This approach is suitable for a controlled environment; e.g. moving person in studio that allows accurate segmentation for 3D object using sparse camera setup (Smolić & Kauff, 2005).

3.1.4 Hybrid-based approach

Several studies have investigated the possibility of deploying depth-based with asymmetric coding approach in order to provide bitrate adaptation (Savas et al., 2012; Gurler & Tekalp, 2013). For coding five views video, asymmetric quality, mixed spatial-resolution, combined asymmetric and multi-view plus depth coding approach are able to reduce bitrate with respect to simulcast video coding by 19.3% to 60%. The graceful bitrate degradation starts by deploying asymmetric quality, asymmetric spatial-resolution then combined asymmetric quality with spatial-resolution. For more bitrate reduction, multi-view plus depth is applied using three views (with their associate depth maps) then two views. Another study has focused on integrating depth-based coding approach with asymmetric chrominance, where multi-view plus depth is applied such that first view contains chrominance information (Shao et al., 2012). At the decoder side, chrominance is reconstructed for the remaining coded views while the intermediate views are synthesised.

The mesh-based coding approach is used to efficiently compress multi-view plus depth (Kim et al., 2007; Merkle et al., 2009a; Keimel et al., 2010). For depth map,

the segmentation using triangles are applied to separate areas that have the same intensity value (mesh triangulation). In this way, the coded depth map would have higher quality than hybrid video codec since the edges are better preserved. This entails improving quality of the synthesised views at the receiver side.

Asymmetric-based coding approach has been integrated with object-based through identifying objects' edges via binary mask (Pinto & Assuncao, 2012; L. & P., 2013). The objective is to code these areas by higher quality than the remaining regions (non-uniform asymmetric quality) that results in improving subjective score than uniform asymmetric quality.

Although the hybrid-based coding approach tends to improve coding performance further than previous coding approaches, it still inherits the challenges from these coding approaches. E.g. accurate depth map is still needed when depth-based coding approach is integrated with other coding approaches.

Since symmetric coding approach has proved its efficiency when multi-view video is coded at low bitrates, it has been used within the investigation reported in this thesis. On the other hand, asymmetric spatial-resolution multi-view video coding provides less encoding, decoding computational complexities and similar subjective quality assessment compared to symmetric coding. Therefore this coding approach is also considered in the investigation reported in this thesis.

The core component of MVC is prediction architecture that distinguishes multi-view video coding from simulcast video coding. It has gained focus from a lot of research in the area of MVC. Therefore, the review will mainly target prediction architectures. The next two sections will focus on symmetric multi-view video coding and mixed spatial-resolution multi-view video coding. In the context of symmetric multi-view video coding, block matching efficiency and prediction architectures are reviewed, while prediction architectures and visual enhancement algorithms are reviewed for mixed spatial-resolution multi-view video coding.

3.2 Symmetric multi-view video coding

This section reviews block matching efficiency, prediction architectures in terms of Reference Frame Selection (RFS) and Reference Frame Ordering (RFO) in addition to coding structures.

3.2.1 Block matching efficiency

The multi-view video coding standard deploys similar inter-picture prediction for motion and disparity estimation. In monoscopic video coding, successive frames are captured via the same view-point. Therefore temporal prediction is efficient, where the variation among predicted and actual block is caused through objects' motion and occluded areas. When similar prediction applies to disparity estimation, the block matching becomes less efficient. The difference in cameras' view angle would affect common spatial information among these frames. The same object would have different spatial information when it is captured via different view-points; this is due to the light illumination, shadow and occlusion. Therefore, it is rationally true that coding efficiency will be marginal when coding sparse located cameras because the amount of common spatial correlation among these cameras is low. In this scenario, simulcast video coding would be preferred than MVC.

Limited work has been conducted to reveal the best coding choice (simulcast video coding or MVC) for the given MVV. Impact of camera separation has been used to provide multi-view videos with either different inter-camera angles or inter-camera distances, where the target is to define the best usage for multi-view video coding. Inter-camera distance refers to the distance among two cameras lenses centres, while inter-camera angle is the angle between two cameras' optical lines. Inter-camera distance and inter-camera angle are used to represent camera separation for linear cameras and convergent cameras setups respectively.

Fecker and Kaup investigated the effect of camera separation on multi-view video coding (Fecker & Kaup, 2005). They used Xmas MVV (101 linear arranged cameras with camera separation of 3 mm), where videos with different inter-camera distances, starting from 3 mm to 90 mm are generated. At all inter-camera distances, MVC provides higher coding efficiency than simulcast video coding. They highlighted that coding efficiency of MVC degrades when camera' separation increases. *Merkle et al.* explored the effect of camera separation in terms of inter-camera distance. They used Rena multi-view video (16 linear arranged cameras with camera separation of 5 cm), where quantisation parameter is varied (Merkle et al., 2007a). They concluded that multi-view video coding becomes efficient when compressing high density cameras at low bitrates. They highlighted that the bitrate per camera saturated at different points depends on quality, where coding multi-view video using low quality settings lead to higher coding gain than coding it at high quality settings. They reported the coding gain of multi-view video coding becomes efficient when both camera separation and the coded videos' quality are decreased. *Abdoli et al.* studied

the effect of the incremental distance between cameras on inter-view prediction (Abdoli et al., 2010). They measured the amount of inter-view prediction using five multi-view videos. The amount of inter-view prediction is 4.4% using the neighbouring cameras, while the amounts of inter-view prediction are 2.6% and 0.53% when inter-camera distance is double twice. They revealed the inverse relationship among inter-camera distance and coding gain. *Bouyagoub et al.* explored the range of inter-camera angles for best usage of stereoscopic video coding for convergent camera setup (Bouyagoub et al., 2010). They reported that stereoscopic video codec should be used rather than simulcast video coding when inter-camera angle among stereoscopic video is up to 20°.

The studies that addressed block matching efficiency either highlighted the relationship among camera separation and coding efficiency of MVC or providing particular threshold for the usage of stereoscopic video coding. Still the criterion for best usage of multi-view video coding is not yet defined.

The next two sections review prediction architectures in terms of reference frame selection and reference frame ordering.

3.2.2 Prediction architectures taxonomy

The majority of prediction architectures focus on reference frame selection. Figure 3-2 shows the taxonomy for deriving prediction architectures. From this Figure, prediction architectures are either derived by block match statistical analysis or proposed heuristically. The first category could be classified into two groups where first group conducts statistical analysis using only frames that belong to time and view directions, while the second group uses frames from all directions. The second category proposes prediction architectures without relying on block match statistics. It would be subdivided into three groups, where the first group proposes prediction architectures that address improving random access. The second group deploys view interpolation prediction that interpolates a subset of frames in order to be used during inter-picture prediction. The third group proposes either single or multiple schemes, where the former proposes prediction architecture, while the latter proposes a set of architectures and evaluates their coding performance.

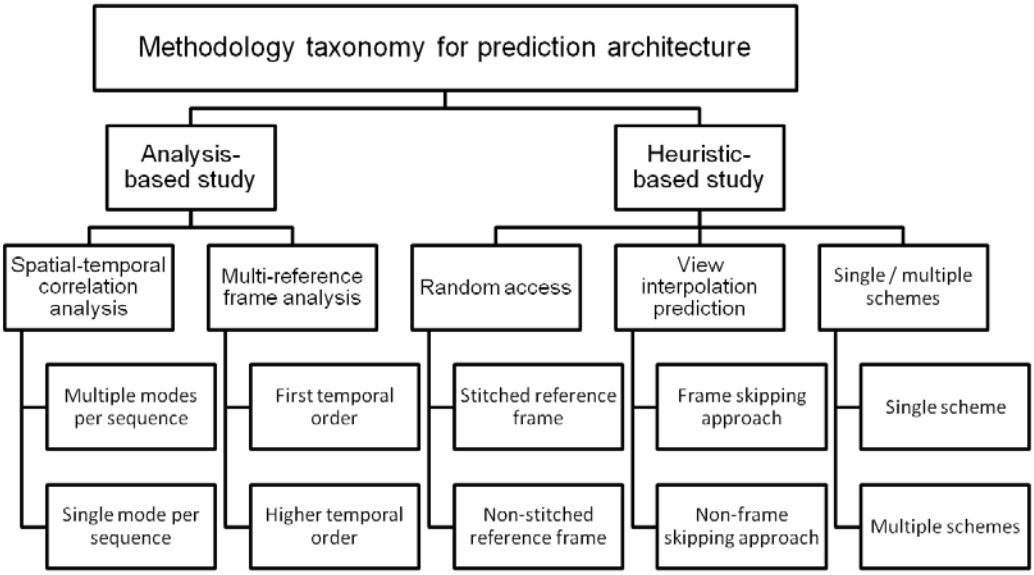


Figure 3-2 Prediction architectures taxonomy for symmetric multi-view video coding

3.2.2.1 Analysis-based study

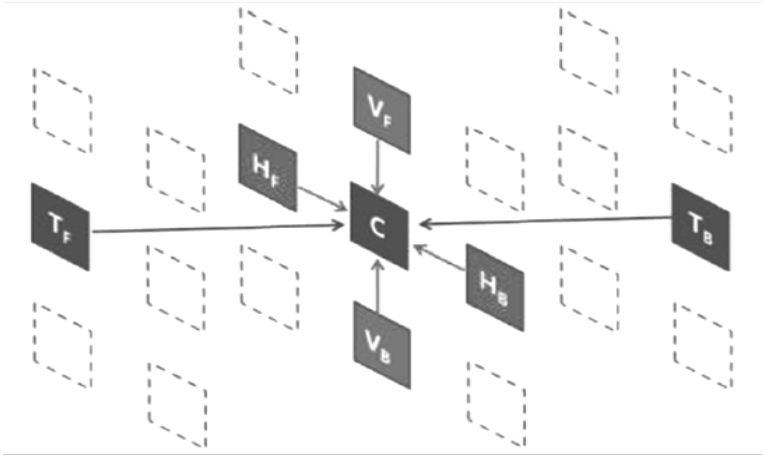
A set of prediction architectures are proposed in the literature based on the analysis of block matching that considers either temporal and spatial reference frames or the entire frames from different directions; temporal, spatial and spatiotemporal reference frames. The former is named as spatial-temporal correlation analysis while the latter is known as multi-reference frame analysis.

3.2.2.1.1 Spatial-temporal correlation analysis

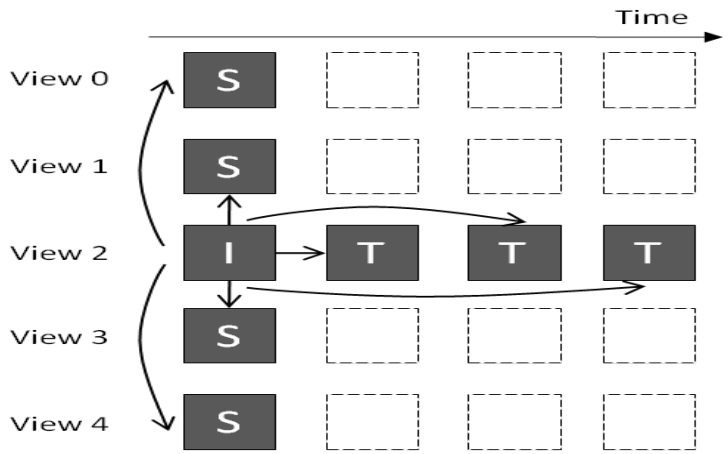
There are two sub-categories that deploy spatial-temporal analysis in order to produce either single or multiple modes per sequence. *Chung et al.* investigated prediction architecture for 2-D camera array (5×9) to derive single mode per sequence (Chung et al., 2008a, 2008b). They analysed block matching for B-frame using backward, forward temporal frames (T_B and T_F), backward, forward spatial frames that belong to neighbouring horizontal and vertical cameras (H_B, H_F, V_B and V_F respectively) as shown in Figure 3-3-a. They reported that temporal prediction provides highest block matching while spatial prediction using vertical cameras gives lowest block matching contribution. They extended the Hierarchical B-Picture (HBP) architecture to cover 2-D camera array.

Other studies focused on different modes per sequence (Zhang et al., 2006, 2009; Lu et al., 2010; Zhang & Cai, 2011). They used typical prediction architecture for H.264/MVC, where the middle view acts as base view. The main idea behind these prediction architectures is to analyse block matching during coding temporal frames

(T frames) that belong to base view (view 2) in addition to neighbouring frames (S frames) that follow anchor frames as shown in Figure 3-3-b. The amounts of temporal and spatial predictions are computed in order to select suitable mode for current GoGOP. *Zhang et al.* proposed in their first study four modes then reduced it to three modes in later studies for GOP which equals twelve (Zhang et al., 2006, 2009, 2011b). *Lu et al.* deployed the same spatial-temporal analysis during switching modes, where GOP equals eight (Lu et al., 2010).



(a)



(b)

Figure 3-3 Spatial-temporal correlation analysis using HBP a) 2-D camera array and b) 1-D camera array (Chung et al., 2008b; Zhang & Cai, 2011)

3.2.2.1.2 Multi-reference frame analysis

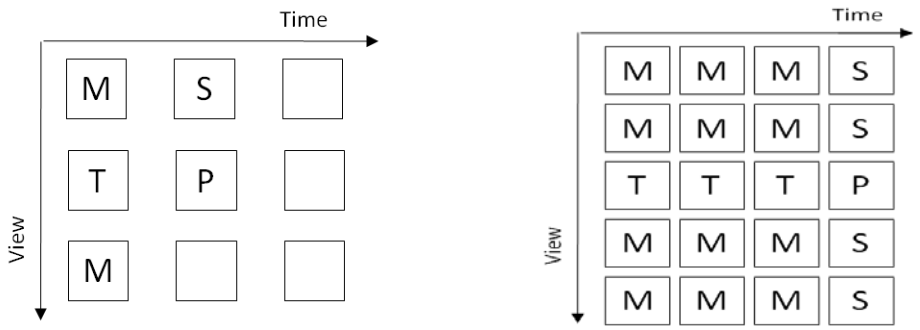
H.264/AVC supports great flexibility for inter-picture prediction through multi-reference frame property. In the context of multi-view video coding, there are three types of frames; temporal, spatial and spatiotemporal frames. Several studies investigated block matching analysis using these frames (Merkle et al., 2006, 2007b, 2007a; Kaup & Fecker, 2006; Yang & He, 2007).

The first temporal order statistical analysis has been applied to reveal which frames provide significant block matching (Merkle et al., 2006, 2007b, 2007a; Yang & He, 2007). *Merkle et al.* analysed block matching among five multi-view videos using one temporal, one spatial and two spatiotemporal frames as shown in Figure 3-4-a (Merkle et al., 2006, 2007a). In their study, intra prediction is disabled while allowing only single coding mode; 16 × 16, with search range adjusted to 32 by 32. The average block matching for temporal, spatial, right spatiotemporal and left spatiotemporal frames are 80%, 9%, 6% and 5% respectively. They tested coding performance when omitting spatiotemporal frames, where average delta Lagrange function is increased by 2.7%. They proposed two categories of prediction architectures that omit spatiotemporal frames. All architectures are based on HBP architecture. The first category deploys inter-view prediction at key frames while the second category uses inter-view prediction at both, key and non-key frames. According to their results, the first category provides coding gain on average 1.6 dB with respect to simulcast video coding while the second category gets additional 0.2 dB gain when inter-view prediction is applied to all frames (Merkle et al., 2007a). *Yang and He* investigated Diagonal Inter-view Prediction (DIP) that provides trade-off among coding efficiency and low delay (Yang & He, 2007). They analysed the amount of inter-view prediction that are exploited through (DIP) and Normal Inter-view Prediction[13] (NIP). The amount of inter-view prediction that came through DIP is in the range 52% to 83 % from the corresponding amount using NIP. They stated that DIP should be used in order to reduce the coding delay.

Higher temporal statistical analysis order has been studied by *Kaup and Fecker* as depicted in Figure 3-4-b (Kaup & Fecker, 2006). They deployed block matching analysis for nine multi-view videos, where 19 reference frames; three temporal, four spatial frames and twelve reference frames are used. They used single coding mode (16 × 16) with limited search area (32 × 32). According to their results, the amount of block matching through spatial and, spatiotemporal reference frames are on average (20% to 30%). They reported a significant amount of block matching when deploying all twelve spatiotemporal frames, however, the role of each frame is small compared to temporal and spatial frames. As a result, they suggested omitting spatiotemporal reference frames, where their prediction architecture involves only three temporal and all spatial frames. They concluded that efficient MVC should use neighbouring three temporal and all spatial frames, where coding gain would be degraded by on average 0.1 dB to 0.2 dB with respect to the same codec that deploys

[13] It uses nearest spatial while DIP uses nearest spatiotemporal reference frame

additional spatiotemporal reference frames (Kaup & Fecker, 2006). They also stated that nearest spatiotemporal reference frames should be selected when the codec allows prediction from this direction.



(a) (b)

Figure 3-4 Block matching analysis using a) Single temporal order and b) Higher temporal order (Merkle et al., 2006; Kaup & Fecker, 2006)

3.2.2.2 Heuristic-based study

The prediction architecture does not necessarily need to be based on block matching analysis as in the previous subsection. Other studies focused on enhancing coding efficiency for H.264/AVC based multi-view video coding using several approaches such as random access, view interpolation prediction and single or multiple schemes.

3.2.2.2.1 Random access

Several studies have proposed prediction architectures that aim to reduce random access (Kimata et al., 2004a; Yebin et al., 2006; Liu et al., 2007; Chen et al., 2008a; Lv, 2013). These studies either deployed stitched reference frame or non-stitched reference frame. The former uses single reference frame to predict frames that belong to the neighbouring views. The latter either uses HBP prediction architecture through modifying GOP structure or uses certain prediction architectures that enhance random access.

Stitched reference frame is an approach that provides efficient random access through relying on higher spatial-resolution monoscopic video (panorama-based)(Li & Ding, 2008; Pourazad et al., 2009b). This video acts as base view, where each view is predicted as an enhancement layer as shown in Figure 3-5. This prediction architecture requires less frame dependency than HBP prediction architecture.

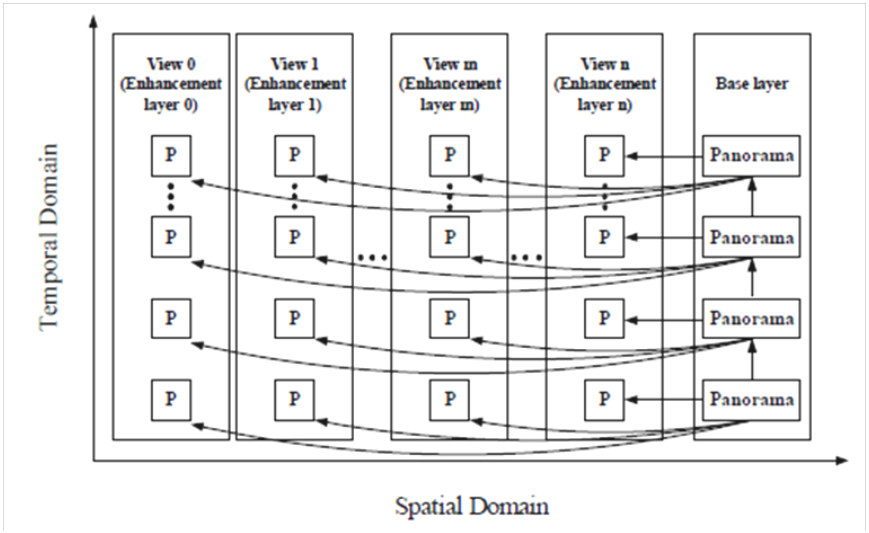


Figure 3-5 Panorama-based prediction architecture (Li & Ding, 2008)

In the context of non-stitched reference frame, several studies have focused on typical prediction architecture, e.g. HBP to improve random access (Park et al., 2008; Lv, 2013; Hussein et al., 2013; Yoon & Kim, 2012). Typical prediction architecture for H.264/MVC uses the first view; V_0 as base view. Random access is improved when the base view becomes the middle view as shown in Figure 3-6 (Park et al., 2008; Lv, 2013). Middle view is identified through global disparity that has lowest disparity among all views (Hussein et al., 2013). Internal configuration for GOP is modified to improve random access through dividing GOP into smaller groups (Yoon & Kim, 2012). *Kimata et al.* proposed prediction architecture based on GOP that contains base-GOP and inter-GOP (Kimata et al., 2004a, 2004b). Frames that belong to base-GOP are predicted using temporal frames while frames that belong to inter-GOP are predicted from frames that belong to the same and different GOP. They proposed Single-Reference and Multiple-Reference prediction architecture (SR and MR) as shown in Figure 3-7. The former omits frames in inter-GOP to be predicted from frames that belong to other inter-GOPs while the latter supports this prediction. *Guo et al.* proposed Global Motion Estimation (GME) that acts as side information when view switching takes place (Guo et al., 2005). Reference frame for switched view is obtained by warping neighbour coded frame using GME. *Kalva and Furht* have used IPPP coding structure with different view arrangement that is inherited from hypercube model (Kalva & Furht, 2005). Eight corners are used to arrange eight-view video. It supports less number of view dependencies (three) than the sequential view prediction structure (seven).

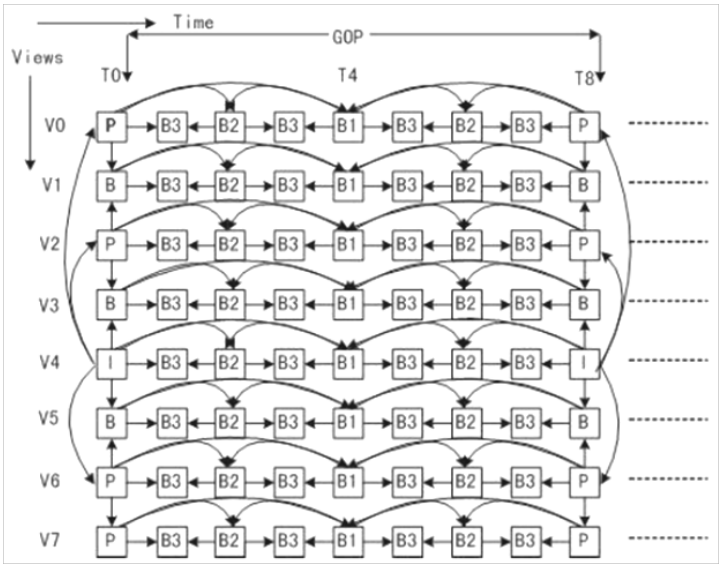


Figure 3-6 HBP prediction architecture using middle view as base view (Lv, 2013)

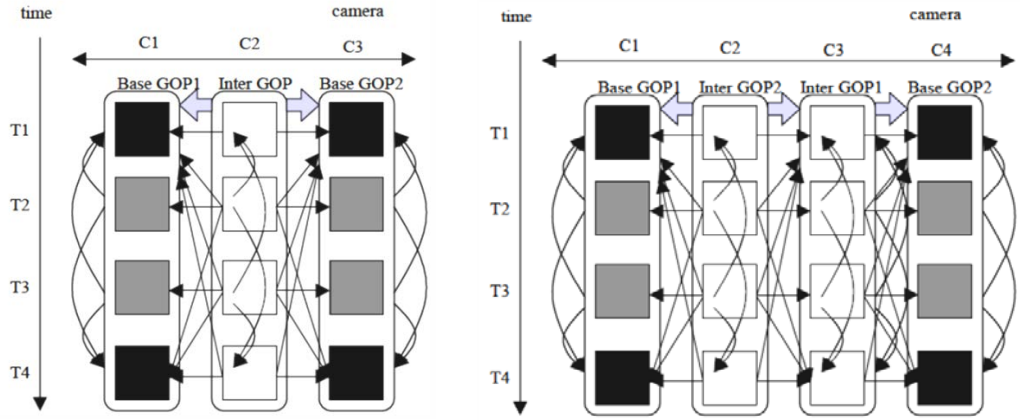


Figure 3-7 GoGOP Prediction architecture, where a) SR and b) MR (Kimata et al., 2004a)

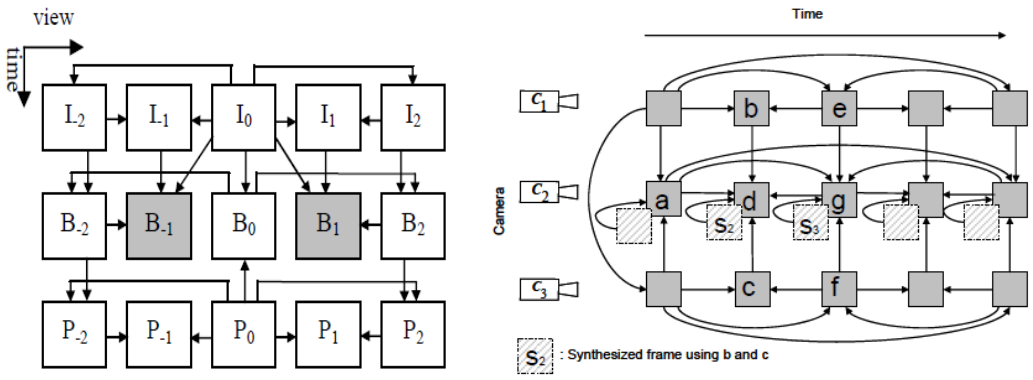
3.2.2.2.2 View interpolation prediction

Reference frame is a crucial part in inter-view prediction, where the reference will be used during disparity estimation and disparity compensation. View interpolation prediction provides another source for inter-picture prediction. There are two categories, where the first is frame skipping approach that omits coding the frame at the encoder side itself, where the synthesised frame will represent the frame at the receiver side. The second is non-frame skipping approach, where the synthesised frame is considered as a potential source for inter-picture prediction.

GoGOP prediction architecture is used to deploy frame skipping approach where part of B-frames are omitted from compression (An et al., 2008). Figure 3-8-a shows the proposed prediction architecture by *An et al.*, where the shaded B-frames are

skipped from coding. *Lee* investigated the possibility of skip coding the frame when camera parameters are known for neighbouring views in addition to coding the difference between original and interpolated frame (Lee, 2013).

In context of non-frame skipping approach, *Kitahara et al.* deployed view interpolation prediction, where camera parameters are known (Kitahara et al., 2006). *Yamamoto et al.* integrated colour correction for all colour channels alongside view interpolation prediction to improve coding efficiency (Yamamoto et al., 2007). *Lee et al.* used disparity estimation to synthesise the given frame. The synthesised frame would be used to predict blocks with different sizes starting from 16×16 to 8×8 (Lee et al., 2007). *Pourazad et al.* proposed prediction architecture that integrates view interpolation prediction with reference frame reordering as shown in Figure 3-8-b (Pourazad et al., 2009a).



(a) (b)

Figure 3-8 Prediction architecture proposed by a) *An et al.* and b) *Pourazad et al.* (An et al., 2008; Pourazad et al., 2009a)

3.2.2.2.3 Single / multiple schemes

Several studies did not follow any of the previous categories. They simply proposed either single or multiple schemes.

A single scheme has been used in few studies, where a novel architecture or frame type is proposed (Oka et al., 2004; Fecker & Kaup, 2005; Oh & Ho, 2007; Flierl et al., 2007). Multi-direction picture (M-picture) has been introduced that supports twenty-one coding modes (Oka et al., 2004). Figure 3-9-a illustrates the prediction architecture that is proposed by *Oka et al.* This frame type has two advantages; it improves inter-picture prediction accuracy and reduces amount of intra-prediction. *Fecker et al.* have proposed transposed picture ordering (Fecker & Kaup, 2005). This coding order starts coding frames that belong to the same time slice together prior to frames that belong to the next time slice. The prediction architecture employs the recent *N+1* frame for inter-picture prediction, where *N* is number of reference frames

as shown in Figure 3-9-b. *Oh and Ho* have presented pyramid GOP structure with flexible search range (Oh & Ho, 2007). I-frame is alternatively used for successive cameras to support low coding delay. *Flierl et al.* have proposed Matrix Of Picture (MOP) that is based on HBP, where it supports view and temporal scalability (Flierl et al., 2007). They integrated histogram matching with their proposed architecture. It improves inter-view prediction through compensating Y, U and V variations among neighbouring views.

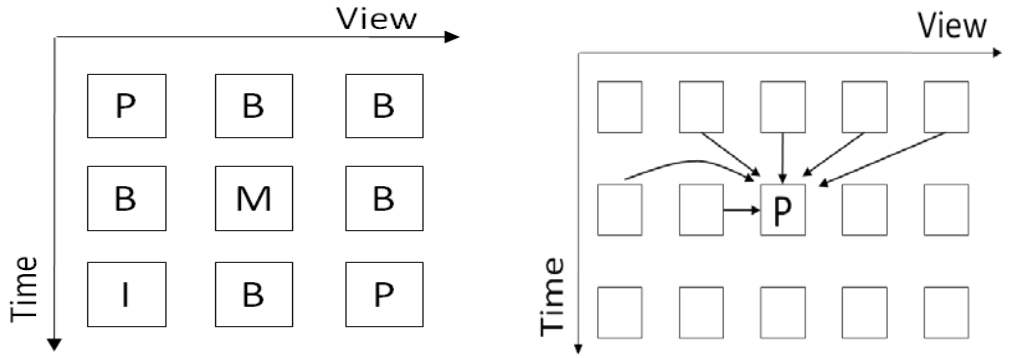
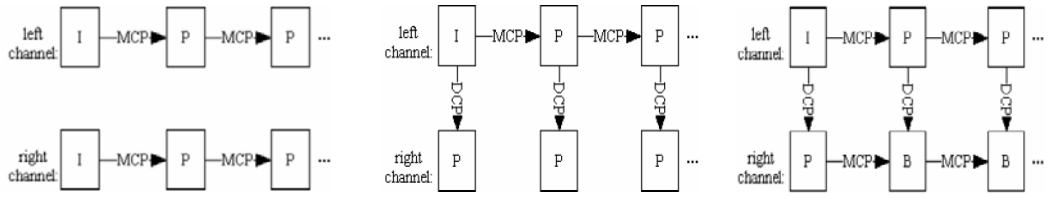


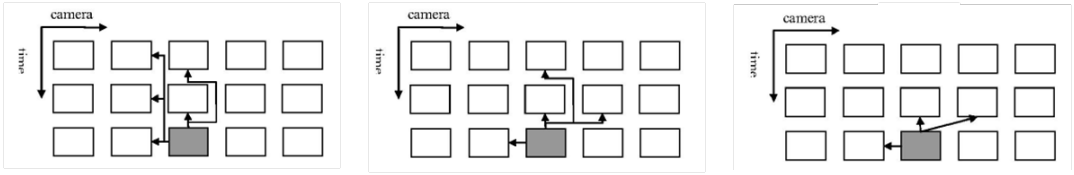
Figure 3-9 Prediction architecture that is proposed by a) *Oka et al.* and b) *Fecker and Kaup* (Oka et al., 2004; Fecker & Kaup, 2005)

Multiple schemes are proposed and evaluated to select the most efficient architecture in terms of coding efficiency (Li et al., 2004; Bilen et al., 2006; Sheikh Akbari et al., 2007). *Li et al.* presented three schemes for stereoscopic video coding as shown in Figure 3-10 (Li et al., 2004). They showed the highest coding superiority when the third scheme is deployed. *Bilen et al.* have compared three schemes (modes) as shown in Figure 3-11. They stated that multi-view video coding is efficient when coding dense camera setup. They showed superior coding efficiency when coding multi-view video that contains scene change by multi-view video coding rather than simulcast video coding. *Sheikh Akbari et al.* have proposed two prediction schemes alongside two reference frame ordering as depicted in Figure 3-12. These reference frame orderings are temporal-first and spatial-first. They compared these modes when coding multi-view videos at different frame rates. They reported that MVC is superior to simulcast video coding when coding MVV at low frame rate. They stated that reference frame ordering has minor effect on the coding performance of MVC.



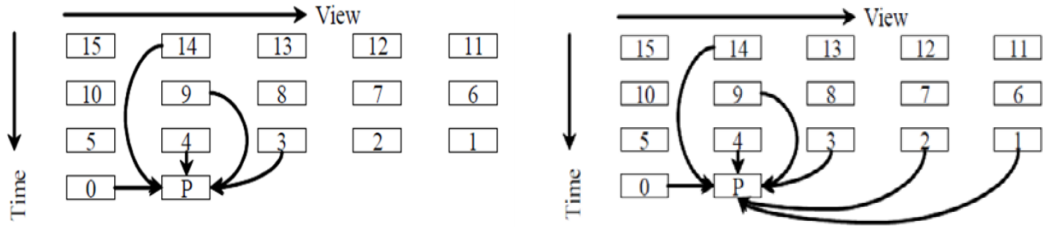
(a) (b) (c)

Figure 3-10 Multiple schemes presented by *Li et al.*, where right frames use a) MCP, b) DCP and c) MCP and DCP (Li et al., 2004)



(a) (b) (c)

Figure 3-11 (a-c) Modes 1, 2 and 3 that are proposed via *Bilen et al.* (Bilen et al., 2006)



(a) (b)

Figure 3-12 (a-b) Modes 1 and 2 that are proposed by *Sheikh Akbari et al.* (Sheikh Akbari et al., 2007)

Several prediction architectures are proposed in the literature. Heuristic-based (non-analysis based) has variety of architectures. Part of these prediction architectures are deployed to improve random access that are suitable for FTV applications. Others rely on view interpolation prediction to improve coding performance of MVC that are suitable for planar camera setup. Other prediction architectures that belong to single / multiple schemes do not provide justification behind their architectures. In the context of analysis-based study, prediction architectures are derived through conducting block matching analysis among reference frames. Spatial-temporal correlation analysis is used to customise HBP architecture according to scene characteristics, where proposed architectures inherit the challenges from HBP architecture that include high computational complexity and memory resources. Few studies have used multi-reference frame analysis that considers all frames on contrary to spatial-temporal analysis, where their studies do not employ all coding modes of H.264 in addition to using limited size of search area. According to the outcomes from these studies, there are no clear clues about

reference frame selection that should be used when H.264/AVC operates at low bitrates. Since multi-reference frame analysis using higher temporal order considers sufficient numbers of frames from all prediction directions, it has been chosen to derive the prediction architectures proposed in this thesis.

Prediction architecture is determined by reference frame selection and reference frame ordering. The majority of studies in the literature have focused on RFS. Few studies have investigated reference frame ordering where they show its importance on coding efficiency of multi-view video coding. The following subsection will therefore review reference frame ordering.

3.2.3 Reference frame ordering

Reference Frame Ordering (RFO) is categorised according to the way it is applied. It is either static or dynamic as shown in Figure 3-13.

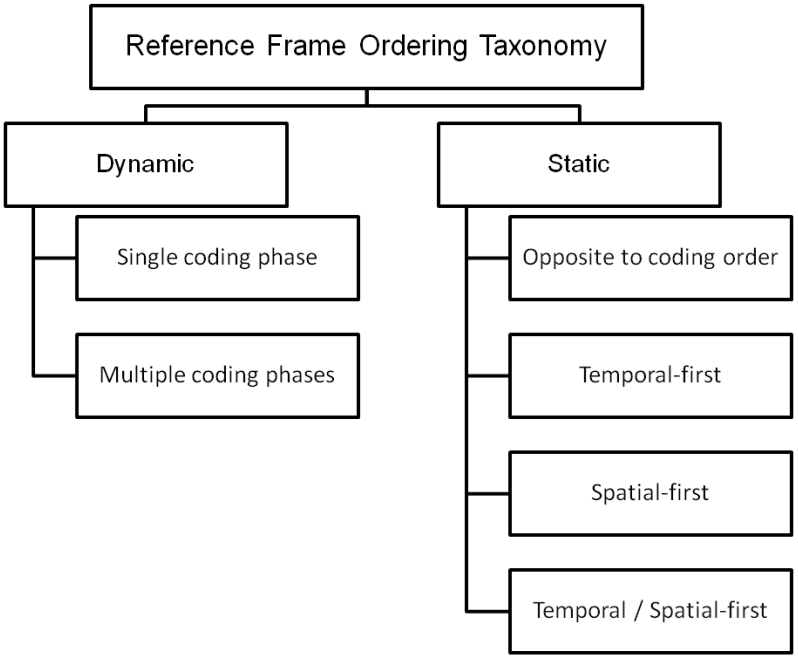


Figure 3-13 Reference frame ordering taxonomy for symmetric MVC

3.2.3.1 Static reference frame ordering

Static reference frame ordering is the common choice for studies that investigate prediction architectures (Fecker & Kaup, 2005; Bilen et al., 2006; Sheikh Akbari et al., 2007). Opposite to coding order is initiated by *Andre* and *Fecker* (Fecker & Kaup, 2005). In fact, this ordering is the normal extension for default RFO in monoscopic video codec. It sorts the reference frames' indices in opposite to their coding order, where the recent coded frame index will be placed first in the List buffer. Since the frames (at the same time slice) that belong to neighbouring views are coded together,

this reference frame ordering assigns the shortest code to nearest spatial reference frame as shown in Figure 3-14-a. *Bilen et al.* have used temporal-first for reference frame ordering (Bilen et al., 2006). They assigned lower indices for temporal reference frames than spatial and spatiotemporal reference frames as shown in Figure 3-14-b. Temporal-first and spatial-first have been deployed separately in two modes by *Sheikh Akbari et al.* as shown in Figure 3-15 (Sheikh Akbari et al., 2007). Temporal-first reference frame ordering places the indices that belong to temporal reference frames prior to the other reference frames. Spatial-first reference frame ordering places the indices that belong to spatial and spatiotemporal reference frames first in the List buffer. Temporal / spatial-first is deployed in multi-view video coding standard (ISO/IEC MPEG & ITU-T VCEG, 2008). The codec controls the reference frame ordering via parameter named *InterPredPicsFirst*. This parameter can select either placing spatial or temporal reference frames first in buffers. Since the majority of the prediction came across temporal direction, the codec uses temporal-first as a default setting. The challenge from using temporal / spatial-first RFO is to determine which source of reference frames is more significant when GOP is large (e.g. GOP size is fifteen).

3.2.3.2 Dynamic reference frame ordering

Few studies have looked into reference frame reordering (Pourazad et al., 2009a; Seungwook & Yang, 2011). *Pourazad et al.* derived reference frame reordering from single coding phase. They proposed histogram-based technique that is used to deploy reference frame reordering alongside view interpolation prediction (Pourazad et al., 2009a). The frequent referral for each reference frame is counted, where the frames' indices are sorted in order to assign the most frequent reference frame a shortest codes. This mechanism improves reference frame ordering for H.264/MVC that uses HBP prediction architecture. *Seungwook* and *Yang* proposed a patent for reference frame reordering, where the suitable RFO is derived from multiple coding phases. They used H.264/AVC based stereoscopic video coding. The algorithm dynamically reorders reference frame indices through coding each frame twice to derive optimum reference frame ordering (Seungwook & Yang, 2011). At the 1st coding cycle, the reference frame order exploited from the previous frame is used while block matching is conducted. In the second coding cycle, the ordering is set according to block matching statistics deployed in the first coding cycle. The numbers of skipped macroblocks in both cycles are compared, where the RFO that leads to higher amount of skipped macroblocks is stored to be used for the following frame.

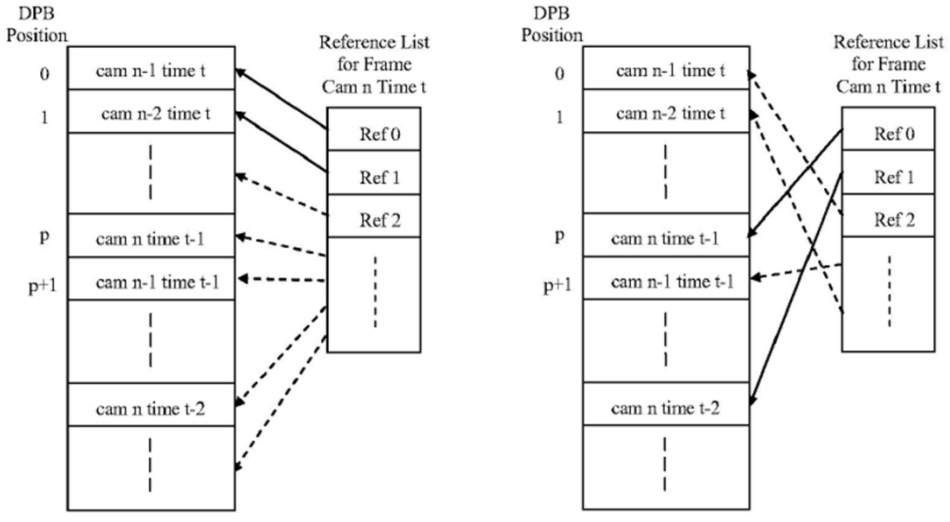


Figure 3-14 Decoded picture buffer with reference frame ordering for a) opposite to coding order and b) temporal-first (Bilen et al., 2006)

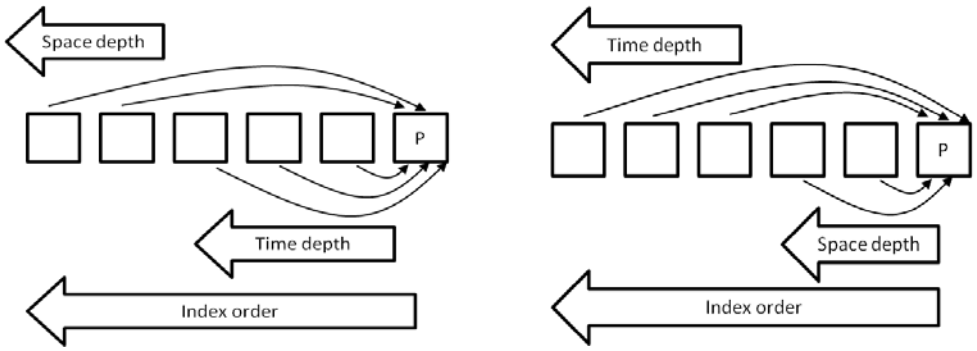


Figure 3-15 Reference frame ordering used by *Sheikh Akbari et al.*, where a) temporal-first and b) spatial-first (Sheikh Akbari et al., 2007)

Prediction architecture either deploys static or dynamic reference frame ordering. The proposed static reference frames ordering in the literature are not theoretically justified. Dynamic reference frame ordering proposed by *Pourazad et al*. has several challenges. First, the encoder needs to signal the reference frame ordering to the decoder when the current order is changed. Secondly, the proposed reference frame reordering does not consider scene change scenarios. For HBP prediction architecture, forward frames are coded first before the frames that belong to previous time slices. Therefore, when a scene changes the frame that belongs to the new scene would be coded and analysed before the frames that belong to previous scene. Since both frames belong to different scenes, the information that is exploited by their algorithm would be irrelevant to the correct reference frame ordering for the current frame. The patent proposed by *Seungwook* and *Yang* gets optimum ordering. Since the proposed algorithm by *Seungwook* and *Yang* encodes frame twice, it does

not fit the requirement of low bitrate applications. After reviewing reference frame ordering, dynamic ordering using single coding phase is investigated in this thesis since it would provide a low computational complexity solution for solving the reference frame ordering.

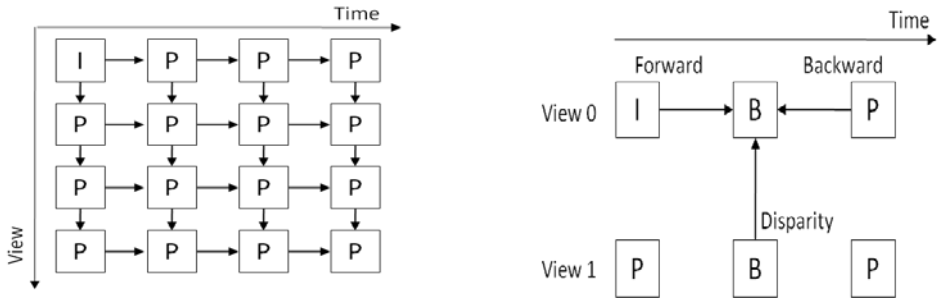
There are two coding structures that are used in the context of multi-view video coding. The following subsection will briefly discuss these coding structures.

3.2.4 Coding structures

Prediction architectures either deploy IPPP or IBBP coding structures as outlined in subsection 2.1.1.4. Different prediction architectures that use these coding structures are compared in terms of coding efficiency, computational complexity and memory consumption (Zhang et al., 2008). Sequential View Prediction Structure (SVPS) using P-frame is a straightforward example for IPPP coding structure that is shown in Figure 3-16-a. They showed that HBP prediction architecture is more coding efficient than SVPS at the expense of higher computational complexity and memory consumption. They measured complexity in terms of minimum number of reference frames that is equal to 58 and 96 for SVPS and HBP respectively. DPB needs to store at least 7 and 21 frames for these architectures respectively. In the context of number of block matches, coding P-frame needs less number of block matches than B-frame. P-frame needs 259 block matches when one reference frame is used. For stereoscopic video coding using HBP (Figure 3-16-b), B-frame needs 160055 block matches (Chiang et al., 2011). It includes forward, backward, disparity, forward plus backward and disparity plus backward. The superior coding performance from using B-frame than P-frame is a result from allowing backward, forward and bi-prediction, on contrary to the latter that uses only forward prediction (Richardson, 2010). This allows several prediction sources when coding B-frame that entails higher prediction accuracy than deploying P-frame at the expense of high computational complexity and memory consumption.

In the context of symmetric MVC, several studies focus on reducing the complexity of HBP in two directions. The first direction is looking into reducing computational complexity, where four different levels are addressed. They are prediction mode, prediction direction, reference frame and block matching (Shen et al., 2010; Zhang et al., 2011a; Khattak et al., 2013). The second direction is reducing memory consumption, where data reuse or parallel architecture are conducted to reduce memory bandwidth (Tsung et al., 2007; Choi et al., 2011; Sampaio et al., 2013). Although studies in both directions achieve significant complexity reduction

compared to HBP architecture, no unified framework has been proposed that reduces both, computational complexity and memory consumption.



(a) (b)

Figure 3-16 a) Sequential view prediction structure using P-frames and b) prediction sources in HBP prediction architecture (Zhang et al., 2008; Chiang et al., 2011)

Prediction architectures could be deployed using either IPPP or IBBP coding structures. Prediction architectures based on IBBP coding structure are more coding efficient than corresponding architectures that deploy IPPP coding structure at the expense of higher computational complexity and memory consumption. Since low bitrate applications prefer coding solution with low complexity and memory requirements, IPPP coding structure is used in the investigations presented in this thesis.

The following section will review prediction architectures and visual enhancement algorithms for mixed spatial-resolution multi-view video coding.

3.3 Mixed spatial-resolution multi-view video coding

This section explores prediction architectures and visual enhancement algorithms for mixed spatial-resolution MVC. Prediction architecture is the core component that distinguishes MVC form simulcast video coding, while visual enhancement addresses improving visual quality for the interpolated frames at the receiver side. The following subsection presents literature review for prediction architectures.

3.3.1 Prediction architectures taxonomy

The taxonomy first classifies prediction architectures in terms of whether the frames arrangement follows conventional mixed spatial-resolution multi-view video or not conventional one. Suppression theory describes the total perceived quality when viewing stereoscopic video that contains views with different qualities. This entails that one of the views has full spatial-resolution frames, while the other has lower spatial-resolution frames. This arrangement is named in this thesis as conventional mixed spatial-resolution that is illustrated in Figure 3-17. The majority of studies

follow conventional mixed spatial-resolution, while few studies proposed other frame arrangements that are referred to as non-conventional mixed spatial-resolution. For each frame arrangement; there are two coding solutions. The first is simulcast video coding while the second is MVC. The latter could be deployed using either Hierarchical B-picture; typical prediction architecture for H.264/MVC or other architectures that are based on IPPP coding structure. Figure 3-18 presents the prediction architectures taxonomy for mixed spatial-resolution multi-view video.

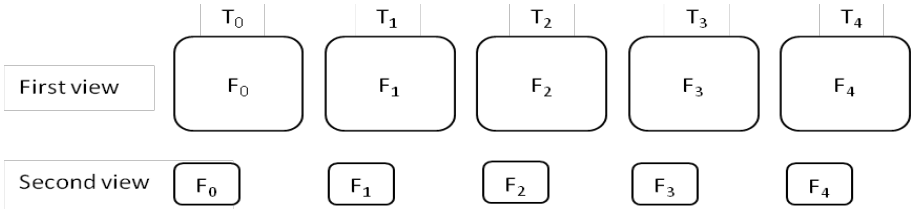


Figure 3-17 Conventional mixed spatial-resolution stereoscopic video

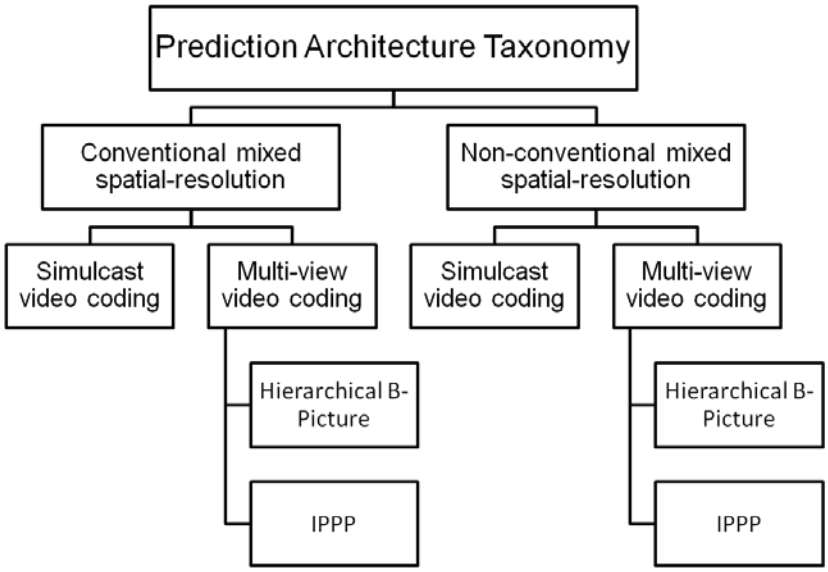


Figure 3-18 Prediction architectures taxonomy for mixed spatial-resolution MVV

3.3.1.1 Conventional mixed spatial-resolution

This subsection covers simulcast and multi-view video coding.

3.3.1.1.1 Simulcast video coding

Simulcast video coding is a common choice for coding conventional mixed spatial-resolution multi-view video coding. Several studies have used simulcast video coding to investigate bitrate allocation, design low pass filters, compare this coding approach with other approaches and to, develop full reference video quality metric.

Mobile 3DTV project is focused on exploring different solutions to support stereoscopic video transmission on Digital Video Broadcast over Handheld devices (DVB) that is supported by the European Union. They studied mixed spatial-resolution stereoscopic video coding using simulcast video coding (Tech et al., 2009b; Brust et al., 2009; Smirnov, 2010). They compared symmetric spatial-resolution with mixed spatial-resolution stereoscopic video coding on small and large displays (Tech et al., 2009a). They reported that 58% and 61% of assessors prefer coded mixed spatial-resolution stereoscopic video for these displays respectively. *Brust et al.* studied optimum bitrate distribution among mixed spatial-resolution stereoscopic video beside complexity analysis with respect to symmetric spatial-resolution stereoscopic video (Brust et al., 2009). They stated that mixed spatial-resolution is suitable for coding stereoscopic video at low bitrate, where coding artefacts are minimised. Optimum bitrate allocation for view with lower spatial-resolution frames is in the range of 30% to 35% from total bitrate. The total complexity for decoding mixed spatial-resolution stereoscopic video is less than decoding symmetric full spatial-resolution stereoscopic video. *Smirnov et al.* compared the coding performance for mixed spatial-resolution stereoscopic video using a set of filters (Smirnov et al., 2010a). The filter groups are standard anti-aliasing filters, standard interpolation filters and Finite Impulse Response (FIR) anti-aliasing filter with variable cut-off frequency (0.1 to 0.9). They showed that variable cut-off frequency gets higher coding performance than other filters.

Other studies focused on comparing asymmetric spatial-resolution with other coding approaches (Bal, 2009; Strohmeier & Tech, 2010; Saygili et al., 2011; Aflaki et al., 2013a). *Bal* has compared subjectively mixed spatial-resolution stereoscopic video with simulcast and multi-view video coding (Bal, 2009). They reported that MVC provides better results than mixed spatial-resolution stereoscopic video. They stated that this coding approach might provide better subjective score when inter-view prediction is enabled at low bitrate. *Strohmeier et al.* compared subjectively several coding approaches that includes simulcast, multi-view video coding, mixed spatial-resolution stereoscopic video coding and video plus depth coding approach (Strohmeier & Tech, 2010). Two coding profiles are used through H.264/AVC; baseline and high profiles at low and high bitrates. At low bitrate, multi-view video coding and video plus depth provide best results among other coding approaches. It is important to note that mixed spatial-resolution stereoscopic video is coded through simulcast video coding for each view. *Saygili et al.* tried to reveal the best coding approach at high and low bitrate through testing asymmetric quality, asymmetric spatial-resolution and symmetric coding (Saygili et al., 2011). They stated that above

threshold[14], asymmetric quality gets highest coding performance in addition to providing fine control for bitrate adaptation while at low bitrate, symmetric coding and mixed spatial-resolution coding are preferable. *Aflaki et al.* compared subjectively symmetric coding, asymmetric quality coding and mixed spatial-resolution (with asymmetric quality) stereoscopic video coding; ∆QP is 2 to 4 at low bitrate (Aflaki et al., 2010, 2013a). They concluded that asymmetric quality with mixed spatial-resolution gets close subjective score to symmetric coding. They highlighted the importance of mixed spatial-resolution in applications that prefers low coding complexity.

Other studies used simulcast video coding to develop quality metric for asymmetric stereoscopic video coding (De Silva et al., 2012). Blurring artefacts that result from the interpolated frames are simulated by Gaussian low pass filter. They reported that HVS has higher degree of tolerance before identifying asymmetric blur than identifying asymmetric quality. This was explained by high frequency that exists in one of the views for mixed spatial-resolution stereoscopic video while blocking artefacts are easily noticed by their additional high frequency.

3.3.1.1.2 Multi-view video coding

Simulcast video coding does not benefit from inter-view correlation that exists in MVV, while MVC exploits visual spatial redundancy among neighbouring views. Prediction architectures are deployed either by IBBP or IPPP coding structure.

Typical prediction architecture for H.264/MVC is deployed for coding mixed spatial-resolution multi-view video as shown in Figure 3-19. Even views have full spatial-resolution, while odd views have low spatial-resolution frames. *Chen et al.* have used this Prediction Architecture (PA) during their studies that aim to reduce decoding complexity for mixed spatial-resolution multi-view video (Chen et al., 2008a, 2008b, 2009a). During disparity compensation, the even or odd samples are extracted directly from decoded reference frame when scaled disparity vector is pointed to integer or half-sample position respectively (Chen et al., 2008a). When scaled disparity vector points to quarter-sample position, closest integer-samples and half-samples are averaged. Therefore, their approach reduces interpolation complexity for generating half-sample which consumes around 40% of the decoding complexity for Hierarchical B-Picture architecture. They improved direct disparity compensation through selecting suitable filter at the encoder side in order to provide accurate sample on the basis of picture and region levels (Chen et al., 2008b,

[14] It is 31 dB and 33 dB for parallax barrier display and full-resolution projection display respectively

2009a). *Quan et al.* integrated asymmetric spatial-resolution as scalability property for enhancement view of stereoscopic video coding (Quan et al., 2011). In symmetric coding, Full spatial-resolution frames that belong to enhancement view are predicted by neighbouring frames that belong to base view in addition to its lower spatial-resolution. The prediction architecture is switched from symmetric to asymmetric video coding through deploying low spatial-resolution frames in enhancement view.

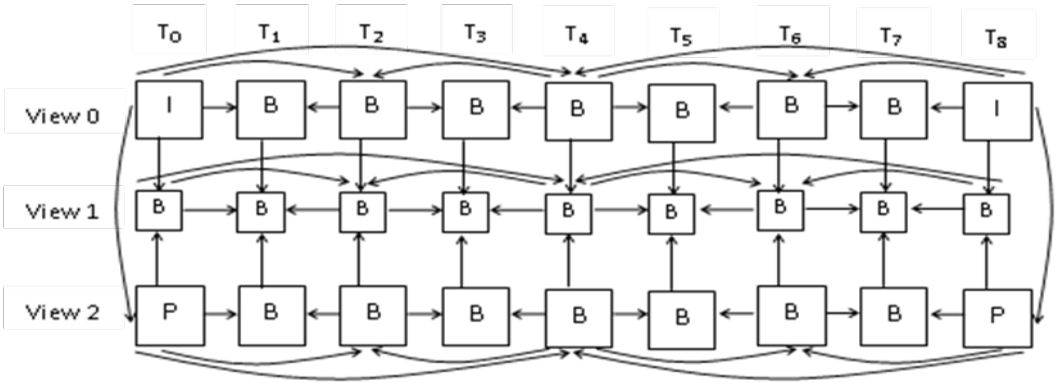


Figure 3-19 HBP prediction architecture for mixed spatial-resolution three-view video

Other studies deployed Hierarchical B-Picture to investigate optimum scaling factor, effect of inter-view prediction direction and examining different decimation methods (Ekmekcioglu et al., 2008b; Tech et al., 2009a; Brust et al., 2010; Aflaki et al., 2013b). *Tech et al.* compared objectively coding performance when coding right view with full spatial-resolution frames alongside left view that are coded using different combinations from filtering, down-sampling and inter-view prediction (Tech et al., 2009a). They showed that enabling inter-view prediction in addition to down-sampling gets highest rate-distortion followed by simulcast video coding for the down-sampled view. Coding both low pass filtered left view with and without inter-view prediction provide lower rate-distortion than former combinations. They highlighted that inter-view prediction for down-sampled left view save 70% of bitrate compared to coding down-sampled left view separately. *Brust et al.* investigated coding performance using different prediction direction for mixed spatial-resolution stereoscopic video coding (Brust et al., 2010). They used asymmetric quality alongside asymmetric spatial-resolution, where full spatial-resolution frames have higher QP than low spatial-resolution frames. They reported that predicting full spatial-resolution by lower spatial-resolution provides equal coding performance when low spatial-resolution frames are predicted by higher spatial-resolution frames at low bitrates. The results revealed by *Brust et al.* are contradicting to the nature of image complexity that is usually increased by decimation (Yu & Winkler, 2013). Based on *Brust et al.* study, the inter-view prediction by FR and LR frames provide

similar inter-view prediction while image complexity is increased by decimation that should have an effect on inter-view prediction (Brust et al., 2010; Yu & Winkler, 2013). *Aflaki et al.* explored different decimation methods for mixed spatial-resolution three-view video coding, where middle view uses full spatial-resolution frames while surrounding views use lower spatial-resolution frames (Aflaki et al., 2013b). They proposed low complexity and high performance decimation methods. Low complexity method down-samples each sample directly that belongs to full spatial-resolution frame without filtering these samples. High performance method decimates integer sample, where remaining sub-pixel samples are generated using corresponding integer and half-samples at low spatial-resolution. They showed superior coding performance when high performance is deployed within mixed spatial-resolution multi-view video coding. *Ekmekcioglu et al.* studied objectively coding performance for mixed spatial-resolution stereoscopic video using different down-sampled scaling factors (Ekmekcioglu et al., 2008b). They deployed several scaling factors starting from 0.3 to 0.9 and compared coding performance for asymmetric with symmetric stereoscopic video coding. They stated that target bitrate affects optimum scaling factor, where highest coding performance at low bitrate is achieved through deploying scaling factor of 0.6 horizontally and vertically.

IPPP coding structure has been deployed in a set of studies to compress mixed spatial-resolution multi-view video coding (Aksay et al., 2006; Fehn et al., 2007; Yang et al., 2009). *Aksay et al.* evaluated asymmetric temporal-resolution and spatial-resolution through seven combinations (Aksay et al., 2006). They compared these asymmetric coding approaches with symmetric stereoscopic video coding. Figure 3-20-a shows prediction architecture that is used to code mixed spatial-resolution stereoscopic video. They concluded that asymmetric spatial-resolution provides optimum solution while asymmetric temporal-resolution is beneficial for slow motion videos. *Fehn et al.* evaluated objectively asymmetric spatial-resolution stereoscopic video coding that is compatible[15] to Digital Video Broadcasting (DVB) (Fehn et al., 2007). They proposed prediction architecture that is named 3D-Digital Multimedia Broadcast (3D-DMB) as shown in Figure 3-20-b. They reported that coding the right view is higher than the corresponding full spatial-resolution frames when coding stereoscopic video at low bitrate. Also, they stated that total bitrate for mixed spatial-resolution stereoscopic video is slightly higher than monoscopic video coding. *Yang et al.* investigated the feasibility of reducing interpolation complexity for decoded mixed spatial-resolution stereoscopic video (Yang et al., 2009). They utilised the

[15] Maximum number of reference frames is three

usage of skip coding mode during disparity compensation, where 3D-DMB prediction architecture is deployed. Skipped macroblocks are directly copied from full spatial-resolution reference frame in order to avoid interpolating these blocks after decoding. They reported that the interpolation complexity is reduced by 30% to 40% of total time consumed by interpolation.

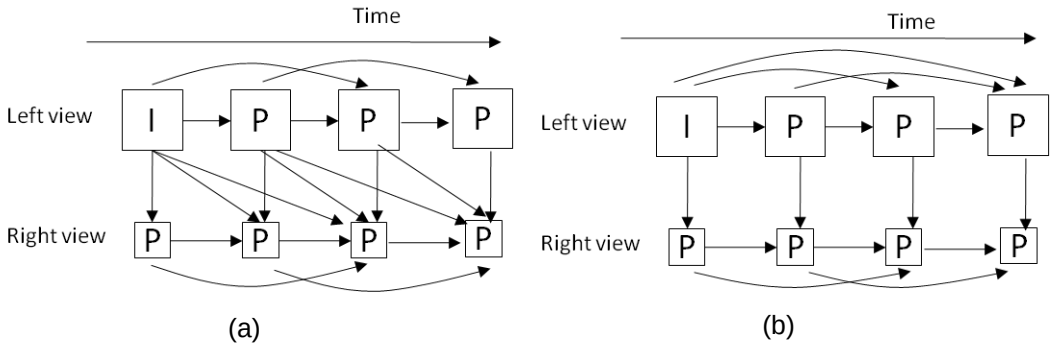


Figure 3-20 Prediction architectures for stereoscopic video coding: a) mode 1 by *Bilen et al.* and b) 3D-DMB by *Fehn et al.* (Bilen et al., 2006; Fehn et al., 2007)

3.3.1.2 Non-conventional mixed spatial-resolution

Some mixed spatial-resolution multi-view video studies do not follow conventional frame arrangement format (Ekmekcioglu et al., 2008a; Yu et al., 2010; Najafi, 2012; Aflaki et al., 2012; Jain et al., 2014). These studies have deployed either simulcast or multi-view video coding.

3.3.1.2.1 Simulcast video coding

Few studies have proposed different asymmetric spatial-resolution frame arrangements formats. *Aflaki et al.* proposed cross asymmetric among stereoscopic video as shown in Figure 3-21 (Aflaki et al., 2012). Decimation is applied differently on both views, where one of the views is horizontally down-sampled while the other view is vertically down-sampled. They used Spatial Index (SI)[16] to measure spatial information for both views in horizontal and vertical directions, where the view with lower SI in horizontal direction is down-sampled in the horizontal direction. They reported that their frame arrangement provide similar results to conventional mixed spatial-resolution format when both are subjectively evaluated on large display (46″ display).

Ankit et al. proposed alternating blur, where asymmetric spatial-resolution is applied in a balanced manner on both stereoscopic views (Jain et al., 2014). Figure

[16] It is computed through extracting edges by deploying SOBEL high pass filter (ITU, 2008).

3-22-b shows alternate blur, where white and grey blocks are sharp and blurred frames respectively. This frame arrangement is different from conventional mixed spatial-resolution frames (single-eye blur) as depicted in Figure 3-22-a. They target reducing eye fatigue that results from watching single-eye blur stereoscopic video by distributing blur level on both views, where their frame arrangement provides better viewing experience for the animated scene than single-eye blur.

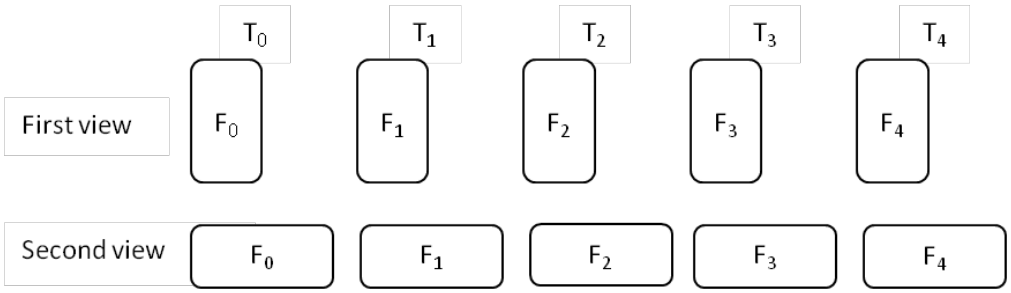


Figure 3-21 Frame arrangement format example by *Aflaki et al* (Aflaki et al., 2012)

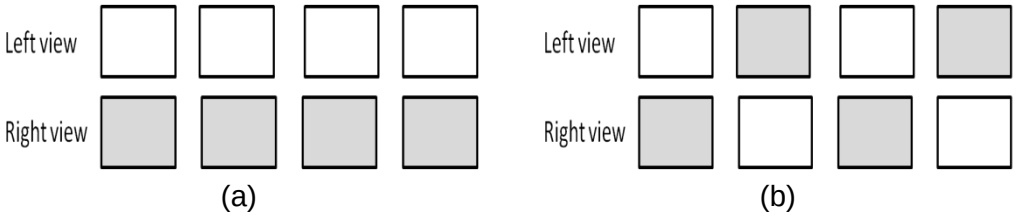


Figure 3-22 Binocular suppression a) single-eye and b) alternating blur (Jain et al., 2014)

3.3.1.2.2 Multi-view video coding

Typical prediction architecture; HBP has been used to compress non-conventional frame arrangements for mixed spatial-resolution multi-view video. *S. Najafi* used low spatial-resolution multi-view video alongside high spatial-resolution still images as shown in Figure 3-23 (Najafi, 2012). They aim to restore high frequency components for coded low spatial-resolution frames by super-resolution technique. The first set (low spatial-resolution MVV) is coded by H.264/MVC as base layer while high spatial-resolution still images are coded as enhancement layer at low frame rate. They show the effectiveness of their algorithm to super-resolve low spatial-resolution frames for scenes that contain fast objects motion.

IPPP coding structure has been used to code anchor frames (Ekmekcioglu et al., 2008a). They used low spatial-resolution frames in majority of anchor frames in order to speed up view switching as shown in Figure 3-24, where GOV is Group Of Views. They stated that their inter-view configurations provide superior view random access

than HBP architecture. *Yu et al.* proposed two sampling directions for frames that belong to the right view as shown in Figure 3-25 (Yu et al., 2010). They deployed 3D-DMB prediction architecture, where I-frame is analysed in terms of Sum of Absolute Transformed Difference (SATD). It is computed based on intra-prediction from either upper or left macroblocks. When SATD for horizontal direction is less than vertical direction, the frames that belong to the right view are horizontally down-sampled; otherwise they are vertically down-sampled.

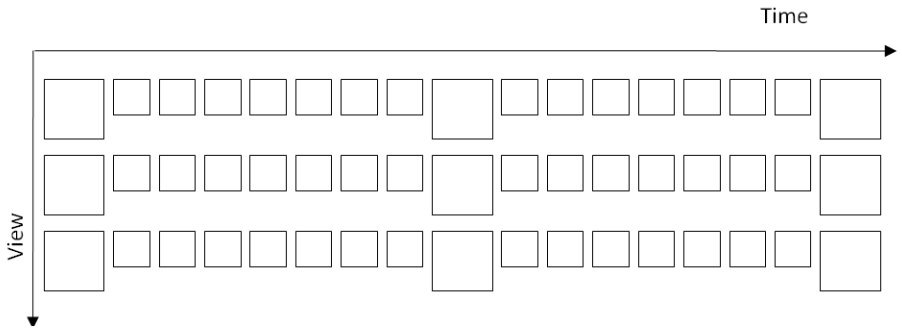


Figure 3-23 Frames arrangement for mixed spatial-resolution MVC (Najafi, 2012)

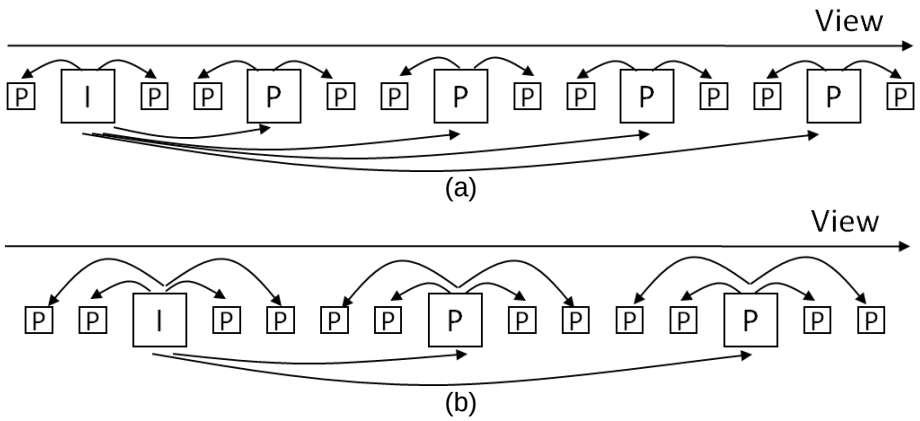
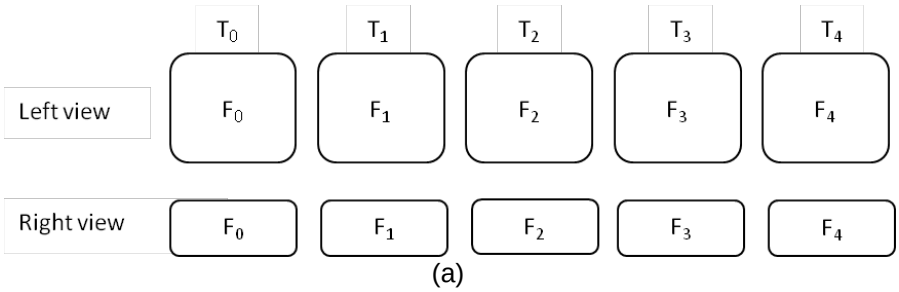
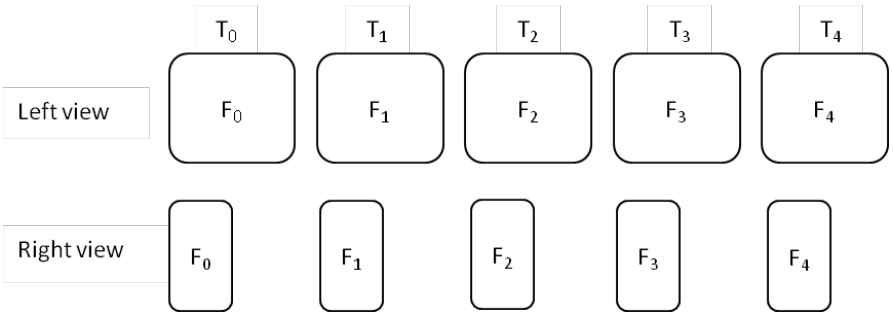


Figure 3-24 Different configurations for inter-view prediction among anchor frames for GOV equal a) 3 and b) 5 (Ekmekcioglu et al., 2008a)





(b)

Figure 3-25 Different down-sampling for right view a) vertical sampling and b) horizontal sampling (Yu et al., 2010)

IPPP coding structure usually needs less computational complexity and memory size compared to IBBP structure. This coding structure is therefore used in the investigation in this thesis for conventional mixed spatial-resolution frames arrangement.

The next subsection will discuss visual enhancement in context of mixed spatial-resolution multi-view video coding.

3.3.2 Visual enhancement algorithms

Suppression theory is used to justify the deployment of mixed spatial-resolution multi-view video coding. It states that HVS would fuse views with different quality, where perceived quality is closer to the view with the higher quality (Bal, 2009; Aflaki et al., 2011). This coding approach needs less coding complexity than symmetric video coding (Brust et al., 2009; Aflaki et al., 2013a). *Michael G. Perkins* has initiated the usage of mixed spatial-resolution stereoscopic video, where disparity-compensated transform-domain predictive coding was applied (Perkins, 1992). They deployed low pass filter to the left view that is sub-sampled by a factor of 4 in horizontal and vertical directions. The subjective score when viewing mixed spatial-resolution stereoscopic video coding is closer to the view that has full spatial-resolution frames (Aflaki et al., 2013a).

Several studies raised several challenges for mixed spatial-resolution stereoscopic video that is either coded by simulcast or stereoscopic video coding. This coding approach (without inter-view prediction) has been compared subjectively to simulcast, stereoscopic video coding (symmetric spatial-resolution) and video plus depth coding approach (Tech et al., 2009b; Strohmeier & Tech, 2010). Mixed spatial-resolution stereoscopic video using simulcast video coding has provided inferior results compared to stereoscopic video coding and video plus depth coding approaches. According to their results, scenes that have slow objects' motion,

complex depth structure or medium spatial information are suited to be coded by asymmetric spatial-resolution. It is worth noting that these studies apply simulcast video coding when assessing this coding approach. *Bal* has stated that subjective assessment for mixed spatial-resolution stereoscopic video is display dependent (Bal, 2009). *Ankit et al.* have raised another challenge due to viewing asymmetric stereoscopic content for a period of ten minutes (Jain et al., 2014). They reported that viewing conventional mixed spatial-resolution stereoscopic video leads to eye fatigue, where most assessors felt eye strain for the eye that receive higher quality. This was the reason behind their proposed frame arrangement to distribute equally the blur on two views that leads to reduce eye strain.

Few studies have focused on enhancing visual quality for coded mixed spatial-resolution stereoscopic video (Tech et al., 2009a; Najafi, 2012). *Tech et al.* applied un-sharp filter to enhance visual quality of coded low spatial-resolution frames (Tech et al., 2009a). They proposed Advanced Mixed-Resolution Stereo Coding (AMRSC), where inter-view prediction and bitrate allocation are investigated. Visual enhancement via un-sharp filter has been subjectively evaluated using two displays; 3.5” and 32”. Although this filter reduces amount of low frequency component and enhances high frequency contents, it magnifies coding artefacts as well. Therefore this filter is not suitable when enhancing coded frames at low bitrates. They stated that full spatial-resolution frame could provide information that can be used during reconstructing low spatial-resolution frame. *S. Najafi* used mixed spatial-resolution for monoscopic and multi-view video coding (Najafi, 2012). The algorithm registers first low spatial-resolution with full spatial-resolution frames. Then up-sampling via non-local means filter followed by de-blurring the up-sampled frame. For multi-view video, they used two types of cameras, first are a set of cameras that capture low spatial-resolution frames at high frame rate while the other set capture high spatial-resolution frames at low frame rate as shown in Figure 3-23. The first set is encoded via H.264/MVC. The up-sampled versions for low spatial-resolution frames are used as predictor for corresponding full spatial-resolution frames. The frame fusion is deployed among temporal frames; therefore the algorithm is affected by scene characteristics, where it becomes less efficient for scenes that have slow objects motion. Their mixed spatial-resolution format is not consistent with conventional mixed spatial-resolution, where low spatial-resolution frames are located between anchor frames. Also the proposed architecture suffers from high complexity at the receiver side.

Although several algorithms have looked into enhancing visual quality for coded low spatial-resolution frames, they do not provide efficient solution in terms of visual

quality or low complexity solution at the decoder side. Therefore, there is a need to find a low computational complexity solution to improve visual quality of the interpolated frames at the receiver side.

In the next section, literature review is summarised alongside a list of studies that are addressed in this thesis.

3.4 Summary

In this section, the literature review is summarised alongside a list of studies that are addressed in this thesis.

3.4.1 Summary of the review

Different coding approaches are used when coding multi-view videos at low bitrates. The depth-based approach has challenges that include depth estimation and view rendering, where an inaccurate depth map and disocclusion would affect the quality of the synthesised view. The main challenge for the object-based coding approach is automatic segmentation that is capable of extracting foreground objects efficiently. Since segmentation is deployed at the sender side, the encoder computational complexity is high. The mesh-based coding approach faces the challenges of mapping geometry models to objects in order to find the best model match for each object, in addition to representing multiple objects in a real scene. The resolution-based coding approach provides a practical solution, where neither depth-map nor segmentation is needed. Symmetric and mixed spatial-resolution coding approaches are more suitable solutions than asymmetric quality when a dependent view is coded at low bitrates. Asymmetric temporal-resolution has shown inferior results with respect to asymmetric quality and asymmetric spatial-resolution coding approaches. Asymmetric spatial-resolution has lower coding complexity than asymmetric quality, since 37.5% of the frames are not coded, when decimation is applied by factor of two horizontally and vertically for stereoscopic video. Therefore, symmetric and, asymmetric (mixed) spatial-resolution multi-view video coding are chosen in the research investigations presented in this thesis.

In the context of symmetric multi-view video coding, block matching efficiency and prediction architectures are reviewed. Several studies have looked into block matching efficiency through investigating the effect of camera separation on the coding performance of multi-view video coding. The target is defining the best usage for multi-view video coding. Part of these studies highlights the relationship among camera separation and the coding efficiency of MVC; others put a hard inter-camera

angle threshold for the best usage of stereoscopic video coding. Still the criterion for the best usage of multi-view video coding is not yet defined.

Several prediction architectures have been proposed in the literature for symmetric MVC. Prediction architectures that are categorised under random access are suitable for certain applications (e.g. FTV), while view interpolation prediction is more suitable for planar camera setup. Single / multiple schemes do not justify the configuration behind their prediction architectures. In the context of an analysis-based study, prediction architectures are justified through analysing block matching statistics among reference frames. Prediction architectures that belong to the category of spatial-temporal correlation analysis inherit the challenges from the HBP architecture that include high computational complexity and memory resources. Multi-reference frame analysis considers all frames, where few studies have used multi-reference frame analysis to derive the configuration of the prediction architecture. These studies do not employ all coding modes of H.264 in addition to using a search area of limited size. According to the outcomes from these studies, there are no clear clues about reference frame selection that should be used when H.264/AVC operates at low bitrates.

Several static reference frame ordering schemes have been proposed that are not theoretically justified. Few studies tried to solve reference frame ordering through proposing algorithms that derive the suitable reference frame ordering dynamically. They provide neither a practical solution that fits requirements of low bitrate applications nor an efficient mechanism that is suitable for videos that contains hard scene changes. There is still a need for an efficient mechanism that is suitable for real time applications and also considers scene change scenario.

Low bitrate applications prefer coding solutions with low computational complexity and memory consumption. Therefore, an IPPP coding structure is used in the investigations presented in this thesis.

In the context of mixed spatial-resolution MVC, prediction architectures and visual enhancement algorithms are reviewed. Non-conventional mixed spatial-resolution MVC tries to provide alternative coding solutions over conventional mixed spatial-resolution MVC. The challenge of non-conventional mixed spatial-resolution is that it does not go through a comprehensive subjective investigation as the conventional frame arrangement. Also, parts of these frame arrangements do not specify how mixed spatial-resolution frames are applied to multi-view video. On the other hand, a conventional frame arrangement has gone through detailed subjective tests in different studies. In addition to this, deploying it into multi-view video is straight forward.

There are two coding solutions for mixed spatial-resolution multi-view video, either each view is coded separately (simulcast video coding) or all views are jointly coded (MVC). Simulcast video coding needs less computational complexity than MVC. Simulcast video coding is not sensitive to camera calibration problem and different lighting conditions. It does not exploit spatial redundancies among neighbouring views, contrary to MVC. Therefore, multi-view video coding is used in the studies presented for mixed spatial-resolution multi-view video.

Hierarchical B-picture and 3D-DMB prediction architectures are used in the majority of mixed spatial-resolution MVC studies. The former architecture inherits the challenges related to significant computational complexity and memory consumption, while the latter architecture justifies neither reference frame selection nor considers reference frame ordering. Therefore, investigating prediction architectures in the context of multi-view video coding that relies on the IPPP coding structure is essential as a potential solution for low bitrate applications.

Although the effect of inter-view prediction direction has been addressed by *Brust et al.*, the results are not consistent with the outcomes reported by *Yu et al.* (Brust et al., 2010; Yu & Winkler, 2013). *Brust et al.* stated that inter-view prediction direction performs equally when the codec that uses either full or low spatial-resolution frames in the base view that operates at low bitrates. *Yu et al.* reported that image complexity is usually increased by decimation that entails affecting the coding efficiency from inter-view prediction. This needs to be addressed to highlight the challenges when inter-view prediction is deployed among mixed spatial-resolution frames. Since there are different decimation and interpolation methods, a comparative study is needed to define potential solutions for these processes.

Few studies revealed negative effects for mixed spatial-resolution stereoscopic video coding. It includes inferior coding results in comparison to the symmetric coding and video plus depth coding approach, when mixed spatial-resolution is simulcast coded in addition to eye fatigue that is reported when watching coded videos. Although there are few studies that target reducing blurriness artefacts via an un-sharp filter and super-resolution technique, they provide neither an efficient solution in terms of visual quality nor a low complexity algorithm. Therefore, a low computational complexity solution is needed to enhance visual quality for the interpolated frames at the receiver side.

The research investigations conducted in this thesis will be presented in the next subsection.

3.4.2 List of studies undertaken

The research investigations are categorised into two phases according to the spatial-resolution that involves symmetric and asymmetric multi-view video coding. The following list outlines the studies toward symmetric multi-view video coding:

- Camera separation is first investigated to determine the best usage for multi-view video coding.
- Prediction architectures have to be investigated, particularly reference frame selection when H.264/AVC operates at low bitrate. Comprehensive statistical analysis of block matching will be used to derive a reference frame selection.
- Reference frame reordering will be investigated in order to efficiently reorder the indices of reference frame dynamically.

In the second part of the thesis, the research focuses on mixed spatial-resolution multi-view video coding, where prediction architectures and visual enhancement for coded low spatial-resolution frames are studied. The following studies outline the investigations undertaken:

- Inter-view prediction direction will be examined for full spatial-resolution and low spatial-resolution reference frames. This needs to be deployed using a symmetric quality configuration among views. This would identify the challenges when predicting frames through different spatial-resolution reference frames.
- Mixed spatial-resolution multi-view video coding implies the need to decimate or interpolate reference frames. Therefore, different decimation and interpolation methods are compared in terms of coding gain and computational complexity.
- Prediction architectures are investigated, where the roles of full spatial-resolution and low spatial-resolution frames need to be explored. The outcomes from the corresponding studies in the first phase would provide clues about reference frame selection and reference frame ordering that facilitates statistical analysis of block matching among mixed spatial-resolution MVC.
- The feasibility for improving the visual quality of the interpolated frames is investigated through exploiting the embedded information in the neighbouring full spatial-resolution frames.

The following chapter targets symmetric multi-view video coding. The first part of the investigation focuses on defining criteria, where multi-view video coding should be used rather than simulcast video coding. Prediction architectures are investigated for stereoscopic and multi-view video coding, using statistical analysis of block

matching. Reference frame reordering is then targeted in order to find a suitable mechanism to reorder the indices of reference frames dynamically.

CHAPTER 4. SYMMETRIC MULTI-VIEW VIDEO CODING

This chapter presents the studies that are relevant to symmetric multi-view video coding. These studies include exploring the impact of camera separation on coding performance of MVC, investigating prediction architectures of H.264/AVC for stereoscopic and MVC in addition to tackling reference frame reordering. Figure 4-1 shows a block diagram for these studies, where the circles labelled by 1, 2 and 3 refer to the impact of camera separation, prediction architectures and reference frame reordering.

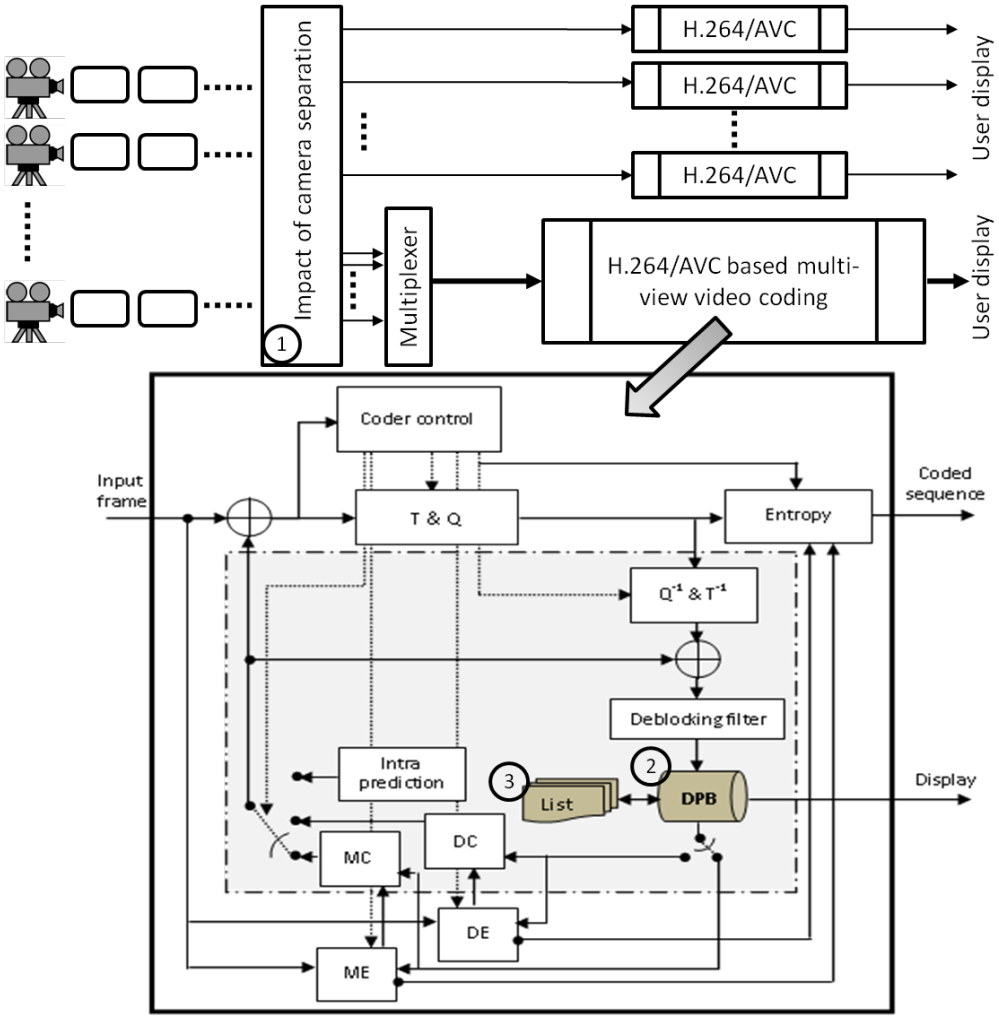


Figure 4-1 Block diagram for the studies conducted in symmetric multi-view video coding

The first study looks into camera separation, where the objective behind it is to determine the best usage for multi-view video coding. Prediction architectures are then studied, in particular reference frame selection that is able to identify reference

frames that have the most block matching contribution during coding of multi-view video at low bitrate. Reference frame reordering is then investigated, where the objective is to find efficient mechanism for dynamically ordering indices of reference frames that would be robust to scene changes.

The following section investigates the first part of the study that explores camera separation effect on the coding performance of multi-view video coding.

4.1 Impact of camera separation on the coding performance of multi-view video coding

4.1.1 Introduction

This section investigates the suitable usage of multi-view video coding through exploring the impact of camera separation on coding performance of MVC. Wide convergent multi-view videos are used in this study since their coding efficiency using multi-view video coding cannot be determined in advance as neighbouring cameras capture the same scene from different angle positions. Camera separation is represented by inter-camera angle that suits coplanar camera setup. This study looks into exploring the range of inter-camera angles, where the multi-view video coding operates efficiently in comparison with simulcast video coding.

4.1.2 Multi-view video with different inter-camera angles

Two datasets have been used throughout this investigation: Break-dancers and Ballet. Both are examples for coplanar camera setup, where they are widely used since their depth-maps are available (Oh et al., 2009; Garcia et al., 2010a). Both were generated via Microsoft research laboratory and these datasets are available from Microsoft research website[17].

Microsoft datasets were generated using eight synchronised cameras that captured fifteen frames per second using PtGrey colour cameras. Each camera generated one-hundred frames. These cameras were positioned in one-dimensional arc configuration with spanning angle of 30°, where each camera had 30° of view with 8 mm lenses. Each frame has an Extended Graphics Array (XGA) resolution and it is represented using RGB colour format (Zitnick et al., 2004).

As this research targets designing multi-view video coding which is suitable for low bitrate applications; the target applications usually prefer display with low spatial-

[17]Online: http://research.microsoft.com/en-us/downloads/5e4675af-03f4-4b16-b3bc-a85c5bafb21d/ /

resolution. Therefore, the original datasets have been spatially decimated to Common Intermediate Format (CIF) resolution.

Generating CIF size frames involves filtering, decimating, cropping and, colour conversion. First, each frame is filtered by 5×5 Kaiser FIR low pass filter in order to reduce the aliasing effect. The filter has cut-off frequency of 0.5 and its coefficients are tabulated in Table 4-1. Filtered frames are spatially down-sampled horizontally and vertically by skipping even samples. The filtered down-sampled frame is cropped starting from point (P_x, P_y) = (120, 47) and (80, 47) for Break-dancers and Ballet respectively. The same starting point is applied to all frames within each dataset to maintain external camera parameters (translation and rotation). The CIF size frame is converted from RGB to YUV colour space. Figures 4-2 and 4-3 show the 1st frame among eight views for Break-dancers and Ballet respectively; these images are scaled down to 10% of its original size. The original, low pass filtered, down-sampled frame and, CIF size frame of Break-dancers are shown in Figures 4-4-a to 4-4-d respectively. The corresponding frames for Ballet dataset are shown in Figures 4-5-a to 4-5-d. The first two images in Figures 4-4 and 4-5 are scaled down to 40% of their original size. The following equations are used during the colour conversion (Ghanbari, 1999);

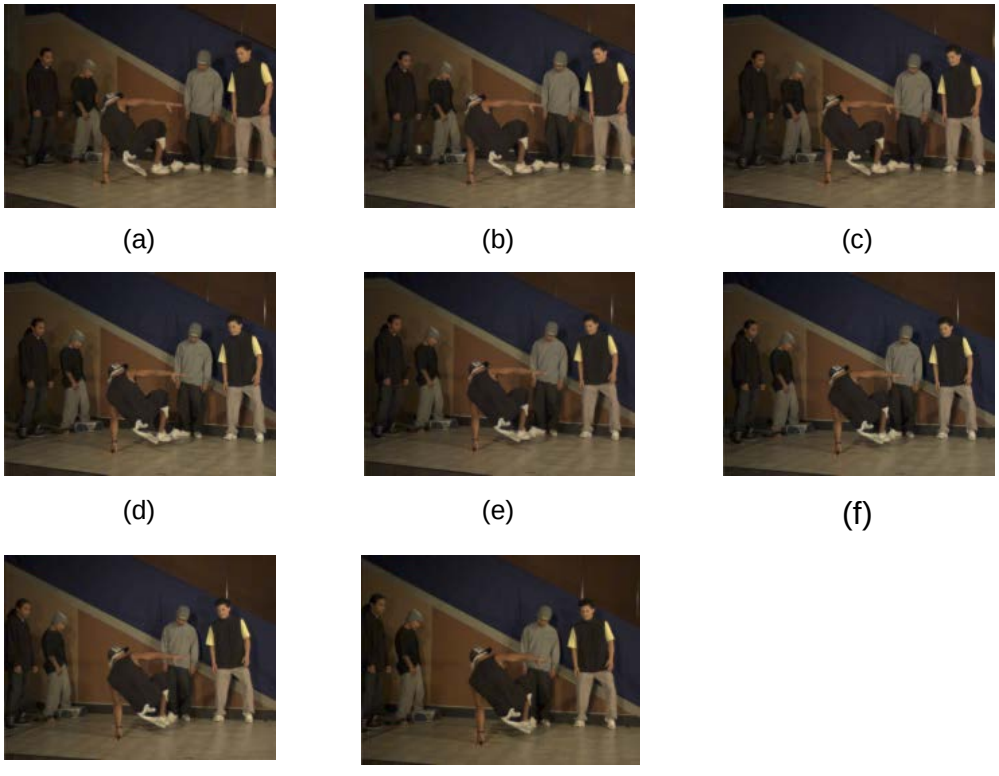
$$Y = 0.2999\,R + 0.587\,G + 0.114\,B \tag{4-1}$$

$$U = -0.148\,R - 0.289\,G + 0.437\,B + 128 \tag{4-2}$$

$$V = 0.615\,R - 0.515\,G + 0.100\,B + 128 \tag{4-3}$$

Table 4-1 Kaiser FIR filter coefficients

0	0	0.0393	0	0
0	0.0653	0.1077	0.0653	0
0.0393	0.1077	0.1511	0.1077	0.0393
0	0.0653	0.1077	0.0653	0
0	0	0.0393	0	0

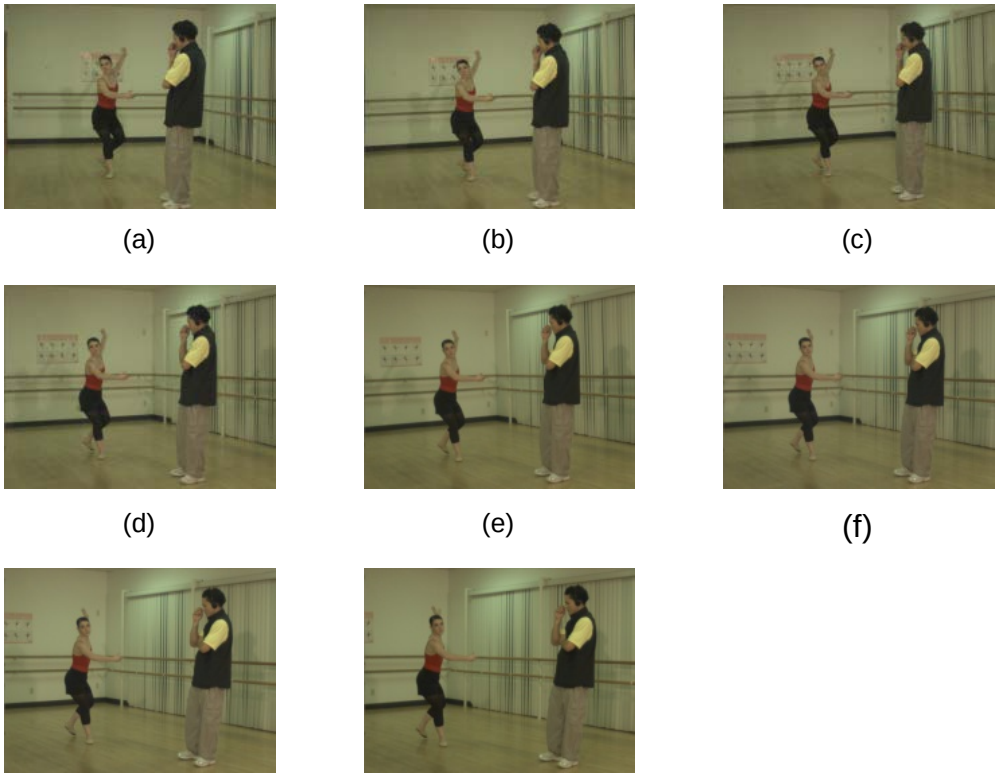


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 4-2 (a-h) show the 1st frame of Break-dancers for camera 0 to camera 7 respectively



(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 4-3 (a-h) show the 1st frame of Ballet for camera 0 to camera 7 respectively

(a)

(b)

(c)

(d)

Figure 4-4 (a-d) show the 1st frame in Break-dancers for camera 0 in its; original, low pass filtered, decimated and, cropped frame respectively

(a)

(b)

(c)

(d)

Figure 4-5 (a-d) show the 1st frame in Ballet for camera 0 in its; original, low pass filtered, decimated and, cropped frame respectively

There are two sets of videos needed to be generated prior to compression. The first set concerns monoscopic videos, each containing single view. Since there are two multi-view videos, each has eight videos; therefore the total number of monoscopic videos is sixteen.

The second set of videos concerns multi-view video. Since multi-view video with different inter-camera angles are required, different videos are generated which contain three different inter-camera angles. First, the inter-camera angles between each camera and reference camera (fifth camera;C_4) is calculated as shown in Table 4-2, where each angle is extracted through panning angles from camera rotation matrices provided in (Zitnick et al., 2004).

Figure 4-6 shows multi-view video with different inter-camera angles, notated by Φ_1, Φ_2, and Φ_3. The camera separation angles; Φ_{iR} and, Φ_{iL} are the angles between centred view to right R and, left L view respectively, i corresponds to certain inter-camera angle, where i = 1, 2 or 3 and the sequence name reflects all the selected camera indices. All possible combinations of multi-view video with different inter-camera angles are then generated for Break-dancers and Ballet as depicted in Tables 4-3 and 4-4. These angles have approximately 4°, 8° and 12° for Φ_1, Φ_2, and Φ_3 respectively.

The selected views are interleaved (multiplexed) into single YUV sequence prior to compression via H.264/AVC based multi-view video coding. The multiplexing starts with a frame from centre view, followed by one frame from each side.

Table 4-2 Camera Separation angles for convergent multi-view videos

Camera Number	Inter-camera angle for Break-dancers (deg)	Inter-camera angle for Ballet (deg)
C_0	-15.8	-18.29
C_1	-12.6	-13.37
C_2	-9.25	-8.41
C_3	-4.63	-4.85
C_4	0	0
C_5	+2.69	+3.46
C_6	+7.52	+8.7
C_7	+10.76	+12.73

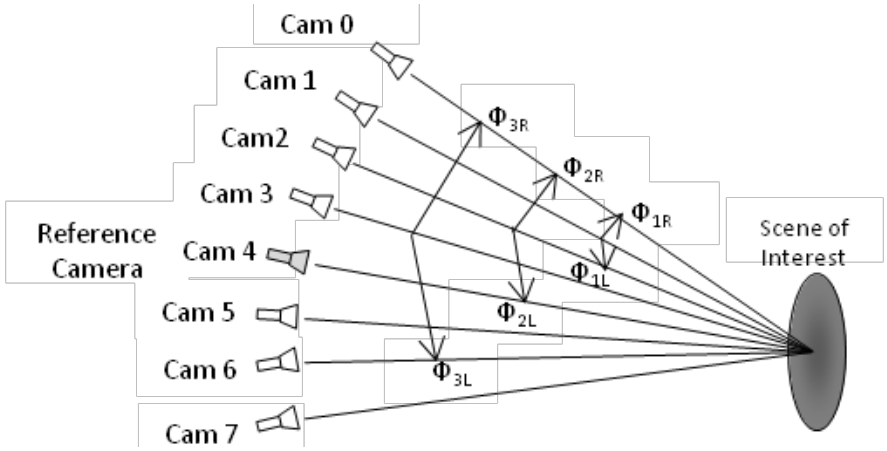


Figure 4-6 Different inter-camera angles for convergent multi-view video

Table 4-3 Inter-camera angles for Break-dancers multi-view video

Φ_i	Φ_1					
Sequence name	**012**	**123**	**234**	**345**	**456**	**567**
Φ_{iR}	3.2	3.35	4.62	4.63	2.69	4.83
Φ_{iL}	3.35	4.62	4.63	2.69	4.83	3.24
Φ_i	Φ_2				Φ_3	
Sequence name	**024**	**135**	**246**	**357**	**036**	**147**
Φ_{iR}	6.55	7.97	9.25	7.32	11.17	12.6
Φ_{iL}	9.25	7.32	7.52	8.07	12.15	10.76

Table 4-4 Inter-camera angles for Ballet multi-view video

Φ_i	Φ_1					
Sequence name	**012**	**123**	**234**	**345**	**456**	**567**
Φ_{iR}	4.92	4.96	3.56	4.85	3.46	5.24
Φ_{iL}	4.96	3.56	4.85	3.46	5.24	4.03
Φ_i	Φ_2				Φ_3	
Sequence name	**024**	**135**	**246**	**357**	**036**	**147**
Φ_{iR}	9.88	8.52	8.41	8.31	13.44	13.37
Φ_{iL}	8.41	8.31	8.7	9.27	13.55	12.73

4.1.3 Experimental setup

H.264/AVC based multi-view video coding has been used via JM software; version 18.0[18] (Sühring, 2011). Recent nine coded frames are used to compress multi-view videos via H.264/AVC based multi-view video coding. Decoded Picture Buffer (DPB) is modified to support this reference frame selection for multi-view video coding. The order of reference frame inside DPB follows default order of JM (opposite to coding order) which sorts the decoded frames in descending order of their coding direction.

[18] Online: http://iphome.hhi.de/suehring/tml/download/

For simulcast video coding, each view is compressed separately using the recent three temporal frames. Search range has been set to cover corresponding points in multi-view videos with different inter-camera angles. It is set to 32, 48 and 64 in the horizontal direction for multi-view video sequences with Φ_1, Φ_2, and Φ_3 inter-camera angles respectively, while search range in the vertical direction is set to 32 among all sequences. The average Peak Signal to Noise Ratio (*PSNR*), of the decoded luminance component for reconstructed videos was calculated at different bitrates. The bitrate starts from 64 Kbps to 1600 Kbps with 128 as delta step size.

4.1.4 Results and discussions

Figure 4-7 and Figure 4-8 present rate-distortion curves when Break-dancers and Ballet videos are coded respectively via MVC and simulcast video coding. It can be seen that coding performance of multi-view video codec outperforms simulcast video codec up to 1.1 dB and 0.3 dB for Break-dancers and Ballet respectively (for Φ_1). Coding multi-view with small inter-camera angle as Φ_1 using multi-view video codec is beneficial. It obtains higher coding efficiency than simulcast video coding for bitrates up to 1088 Kbps and 320 Kbps for Break-dancers and Ballet respectively. Coding performance is decreased when inter-camera angle increases for both multi-view video datasets. For high inter-camera angle such as Φ_3, MVC is beneficial for bitrates up to 576 Kbps for Break-dancers, while it obtains inferior results compared to simulcast video coding for Ballet. Therefore, multi-view video coding is used for Break-dancers dataset with inter-camera angle up to 12°, while it is used for Ballet when corresponding angle is 4°.

It is clear that using multi-view video coding for Break-dancers brings significant coding performance with respect to Ballet dataset. Based on these results, the multi-view video coding efficiency is not only dependant on the target bitrate and inter-camera angle but also on scene complexity. To clarify this, Temporal Index (TI)[19] is used among temporal and spatial frames for both datasets (ITU, 2008). Figure 4-9 shows temporal index, where X-axis and Y-axis are frame numbers and their TI respectively. TI curves using both temporal and spatial frames for Break-dancers are crossed over while TI curve using temporal frames is lower than the corresponding curve for spatial frames using Ballet. The average temporal index using temporal and spatial frames are 14.3 and 15 for Break-dancers, while 8.3 and 21.2 are the corresponding values for Ballet respectively. From these Figures, Break-dancers

[19] It measures amount of motion difference among successive frames (ITU, 2008)

multi-view video has balanced amount of correlations among temporal and spatial frames while Ballet has dominant temporal correlation.

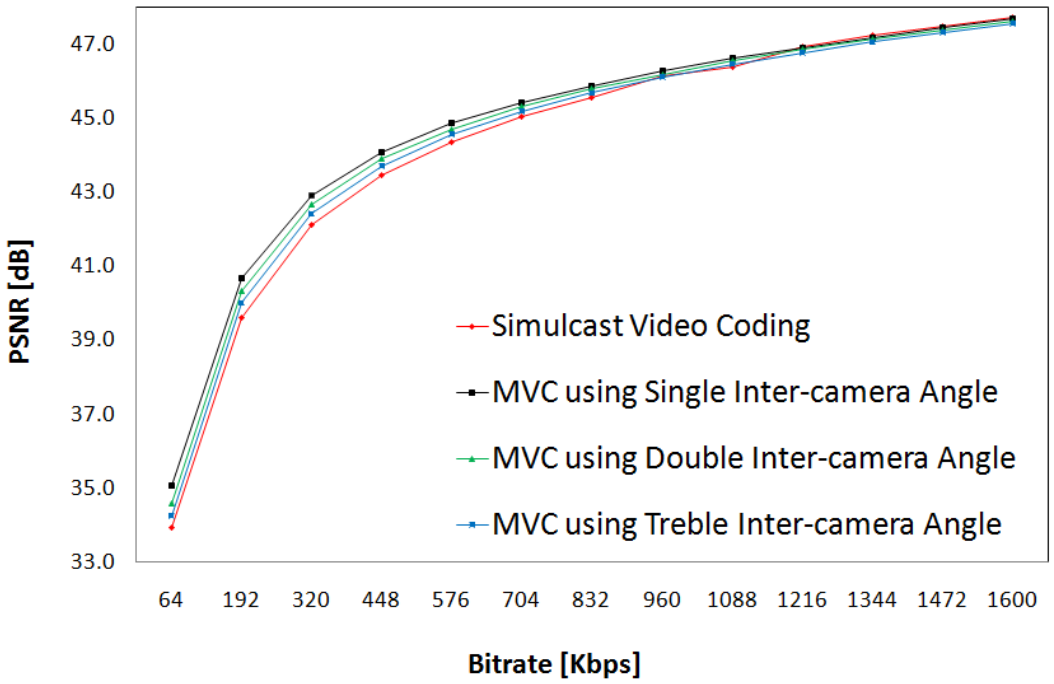


Figure 4-7 Rate-distortion curves for coding Break-dancers videos

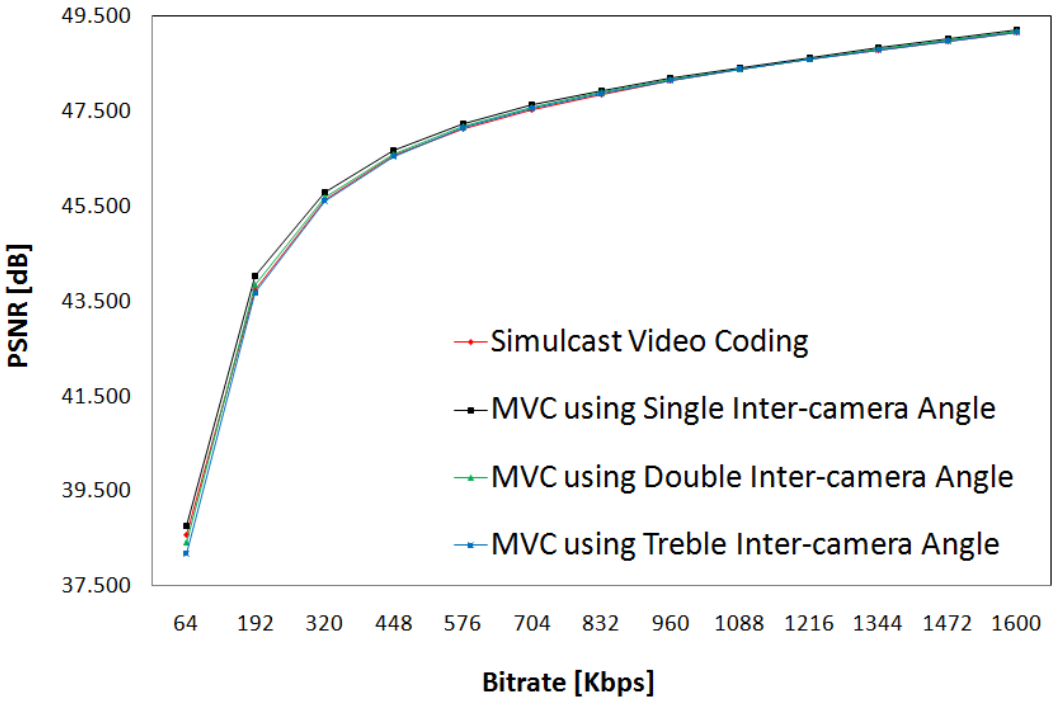
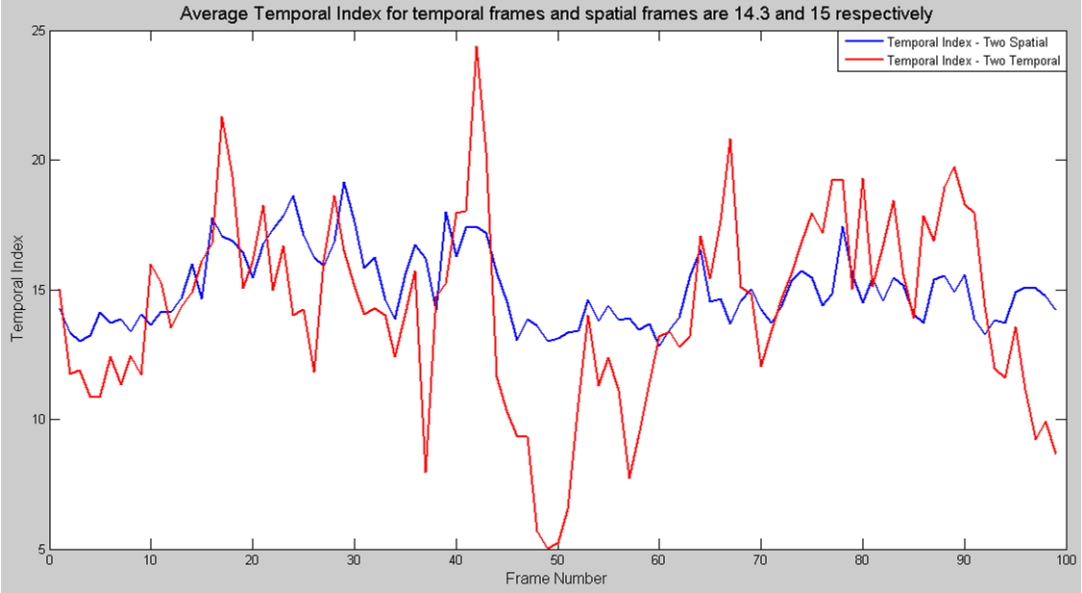
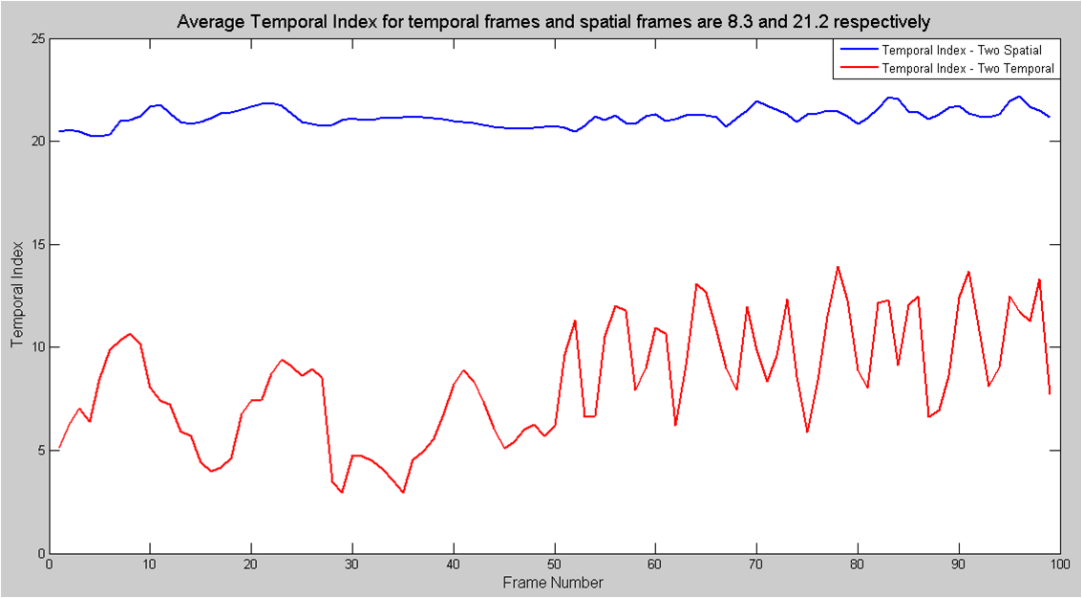


Figure 4-8 Rate-distortion curves for coding Ballet videos



(a)



(b)

Figure 4-9 TI among temporal and spatial frames for a) Break-dancers and b) Ballet

4.1.5 Conclusions

In this section, the impact of camera separation on the coding performance of MVC is investigated for wide-baseline cameras, where inter-camera angle is used to define a criterion for suitable use of MVC. From the results, the suitable usage for MVC depends on the amount of temporal correlation exist in MVV, where a dataset with dominant temporal correlations has lower inter-camera angle threshold (4°) than a dataset with balance temporal and spatial correlations (12°). This entails that inter-camera angle is not a sufficient criterion to decide the best coding solution for the

given multi-view video. In fact, scene characteristic plays a more significant role than camera separation, where objects' motion and scene complexity affect the amount of blocks that are predicted through spatial frames.

The following section will explore prediction architectures for H.264/AVC based stereoscopic video coding (simplest case for multi-view video coding).

4.2 Stereoscopic video coding using statistics of block matching

4.2.1. Introduction

In this section, prediction architectures for H.264/AVC based stereoscopic video coding are investigated at low bitrate. Quantitative statistics method for H.264/AVC based stereoscopic video coding is used to derive RFS and RFO. In the following sections, the generated stereoscopic videos are introduced then statistical analysis of block matching is conducted. The proposed prediction architecture is validated through its coding performance among other prediction architectures and the last subsection concludes the outcome by this investigation.

4.2.2. Stereoscopic videos generation

Microsoft multi-view video (Break-dancers) has been used which is outlined in section 4.1.2. Two additional multi-view videos are considered, Race1 and Exit datasets that have different scene characteristics. Race1 dataset has fast global motion while Exit dataset large disparity with slow objects' motion (Zhang et al., 2011a). Both are generated through capturing the scenes using eight cameras that are placed in linear setup. Each view is stored in YUV 4:2:0 format, where its spatial-resolution is Video Graphics Array (VGA). Race1 MVV is provided via KDDI (available online[20]), where its camera separation is 20 cm and captures 30 FPS. Mitsubishi Electric Research Laboratories, MERL provided Exit MVV (available online[21]), where its cameras capture 25 FPS and their camera spacing is 19.5 cm (Zhang et al., 2011a).

Race1 and Exit datasets are decimated, where their luminance components of each frame are low pass filtered via Kaiser FIR filter and spatially down-sampled to Quarter Video Graphics Array (QVGA) resolution size.

[20] online: www.mmnt.net/db/0/0/ftp.ne.jp/040/KDDI/multiview/Race1
[21] online: ftp.merl.com/pub/avetro/mvc-testseq/orig-yuv/exit/

Stereoscopic videos are generated from Break-dancers, Race1 and Exit such that seven stereoscopic videos are generated from each MVV. Figure 4-10 shows the frames interleaving for stereoscopic video, where each block represent a frame.

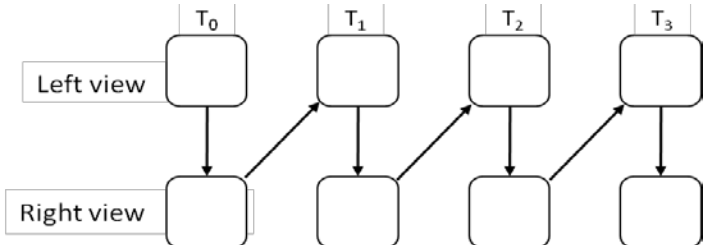


Figure 4-10 Multiplexing frames generated from both cameras into single sequence

4.2.3. Statistical analysis of block matching among reference frames

A statistical analysis is applied so that all coding modes, intra-prediction and rate control are enabled. Two bitrates are considered; 64 Kbps and 192 Kbps which reflect coding each video at low and medium bitrate respectively. Recent seven frames are included in the prediction architecture as shown in Figure 4-11, where T_0-T_2 and M_0-M_3 are the temporal and spatiotemporal frames respectively and *S* is the spatial frame. These frames are sorted in the descending order as depicted in Figure 4-12.

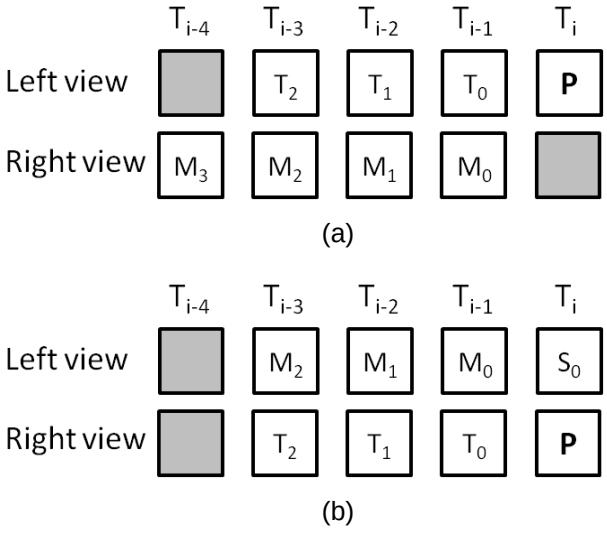


Figure 4-11 Block diagram of reference frames used in the statistical analysis for a) left view and b) right view

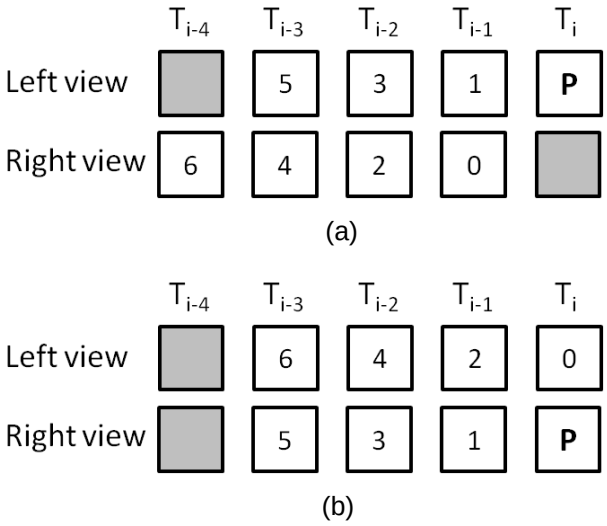


Figure 4-12 Reference frame ordering for frames in a) left view and b) right view

Table 4-5 shows the results for the Break-dancer's statistical analysis of block matching at 64 Kbps and 192 Kbps. These results represent the average statistics of coding seven pairs of the stereoscopic adjacent views. From Table 4-5, it can be seen that the neighbouring reference frames (nearest temporal and spatial) have significant contribution for block matching. Also, from this table, it is obvious that the distribution of the block matching amongst reference frames is inconsistent with the position of the reference frames indices in the buffer List 0. Reference frame T_0 is used for predicting the majority of blocks for right and left views at 192 kbps and left view at 64 kbps.

Table 4-5 Statistics of block matching amongst reference frames for Break-dancers using the descending order frame indexing at bitrate a) 64 kbps and, b) 192 kbps

REF	T_0	T_1	T_2	S_0	M_0	M_1	M_2	M_3
Left	48.23	4.2	3	n/a	40.3	2.5	0.82	0.95
Right	34.63	2.4	1.7	58.7	1.8	0.45	0.35	n/a

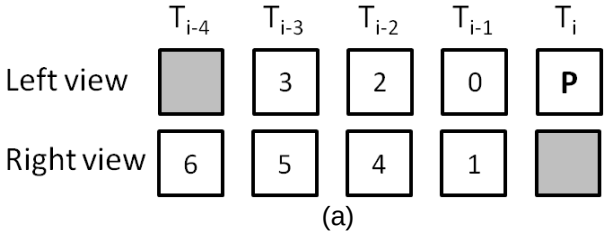
(a)

REF	T_0	T_1	T_2	S_0	M_0	M_1	M_2	M_3
Left	70.4	7.5	4.8	n/a	13.46	2	0.84	1
Right	54	5.1	3.1	35.31	1.6	0.5	0.39	n/a

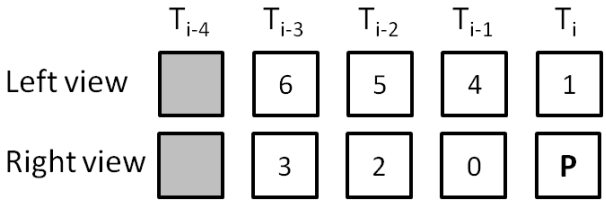
(b)

Based on the previous results, additional coding performance would be obtained through placing the reference frames appropriately based on their block matching contributions. Therefore, the reference frames are first indexed according to their contributions in block matching using the resulting statistics from the first set of experiments beside their spatial position to the current frame, as shown in Figure 4-

13. Another statistical analysis of block matching is performed using the proposed reference frame indexing. The results for Break-dancers are tabulated in Table 4-6. It can be seen that the temporal frames have higher contribution in prediction than the spatial frames. It entails that coding performance of the stereoscopic video codec could be increased by using the proposed reference frame indexing.



(a)



(b)

Figure 4-13 Reference frame order (according to their block matching contribution among reference frames for coding frames a) left view and b) right view

Table 4-6 Statistics of block matching amongst reference frames for Break-dancers using the proposed frame indexing order at bitrate a) 64 Kbps and, b) 192 Kbps

REF	T_0	T_1	T_2	S_0	M_0	M_1	M_2	M_3
Left	85.98	4.4	1.92	n/a	5.59	0.78	0.6	0.73
Right	72	2.96	1.16	22.57	0.35	0.68	0.28	n/a

(a)

REF	T_0	T_1	T_2	S_0	M_0	M_1	M_2	M_3
Left	81.47	7.22	3.28	n/a	5.54	0.93	0.66	0.9
Right	66.4	5	2	25	0.9	0.4	0.3	n/a

(b)

The contribution of the nearest temporal reference frame is increased in the proposed reference frame order rather than previous RFO. From Table 4-6, there is a relationship between target bitrate and inter-picture prediction. Nearest temporal reference frame; T_0 has the highest contribution of block matching, however, the percentage compared to the remaining reference frames decrease with the growth of bitrate. The contribution of the other temporal frames; T_1 and T_2 increase proportionally with the target bitrate.

4.2.4. Proposed prediction architecture

Prediction architecture has been proposed based on the previous block matching analysis. Although the results reported in the previous subsection are based on block match statistics for one dataset (Break-dancers), it provides a reasonable way to reveal the reference frame contribution in terms of block matching since the dataset has balanced amounts of temporal and spatial correlation. Four reference frames are included in the proposed PA, they are T_0, T_1, T_2 and, S_0 or M_0[22]. They have more than 97.5% of block matching of P-frames when coding Break-dancers at 64 Kbps and 192 Kbps. Figure 4-14 shows the proposed prediction architecture, for coding stereoscopic videos, where the number in each block represents reference frame index inside List 0.

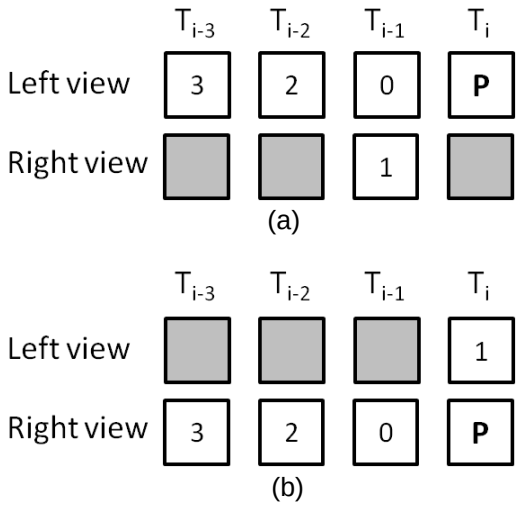


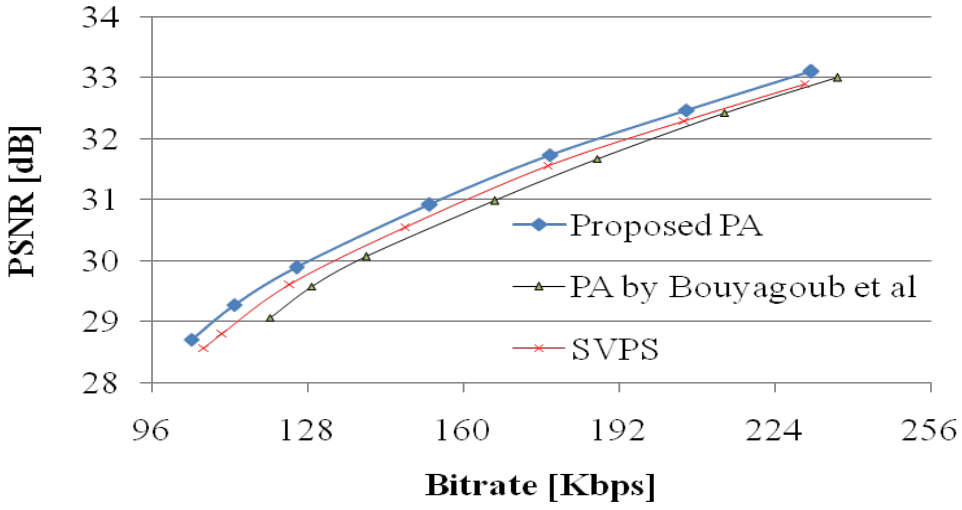
Figure 4-14 Block diagram of the proposed prediction architecture for coding a) left view and b) right view

4.2.5. Results and discussions

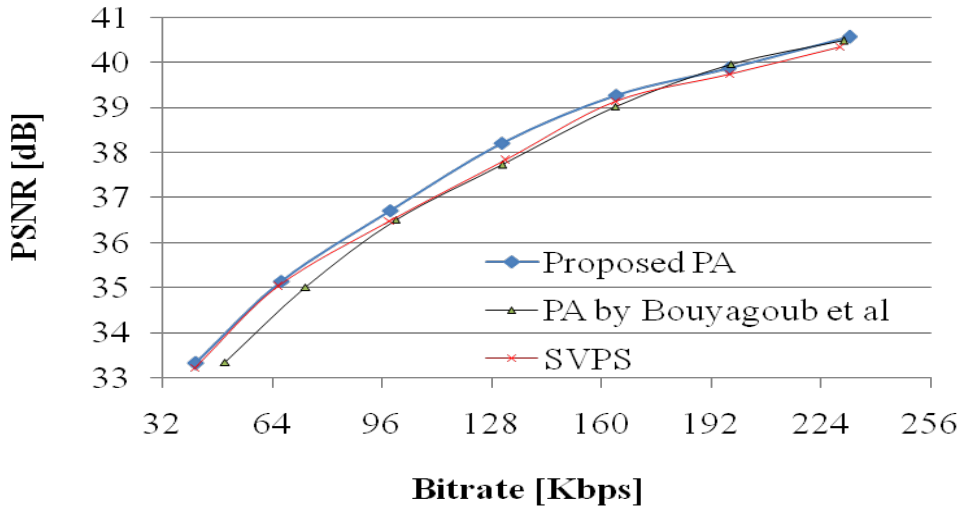
The proposed prediction architecture is evaluated using videos from views number 3 and 4 of Race1 and Exit datasets. The former dataset has global fast objects motion while second dataset has low objects motion and large disparities (Zhang et al., 2011a; Khattak et al., 2013). These videos are coded using H.264/AVC based stereoscopic video codec. The same sequences are coded via two prediction architectures. The first Prediction Architecture (PA) is Sequential View Prediction Structure (SVPS) (Zhang et al., 2008). The second prediction architecture is the one proposed by *Bouyagoub et al.* (Bouyagoub et al., 2010). The first prediction architecture relies on nearest temporal and spatial reference frames while second is

[22] The last reference frame for right or left view

most recent PA amongst IPPP coding structures for stereoscopic video coding. Figure 4-15 shows the *PSNR* results for Race1 and Exit videos. From these results, it can be seen that the application of the proposed prediction architecture improves the coding gain of the H.264/AVC compared to the same codec that uses PA presented by *Bouyagoub et al.* by up to 0.37 dB. The proposed PA improves coding gain of the H.264/AVC to same codec that deploys SVPS by up to 0.49 dB.



(a)



(b)

Figure 4-15 Coding performance of the stereoscopic video codec using the proposed prediction architecture among other prediction architectures for a) Race1 and b) Exit

4.2.6. Conclusions

The Codec that deploys the proposed prediction architecture is found to give better coding performance than the same codec that uses the other two prediction architectures. By exploiting the information derived from block matching statistics,

reference frame ordering is adjusted in a way to be coherent with the role among reference frames in terms of inter-picture prediction that entails improving coding performance of the codec. The proposed PA outperforms a set of prediction architectures that use IPPP coding structure by coding gain up to 0.49 dB.

The following section will investigate prediction architectures for H.264/AVC based multi-view video coding through analysing block matching statistics among neighbouring frames.

4.3 Multi-view videos coding using statistics of block matching

4.3.1 Introduction

This section studies H.264/AVC prediction architectures via block matching analysis. The philosophy of applying statistical analysis to define reference frames that have significant role in block matching while identifying their indices' order that are consistent with their block matching contributions.

4.3.2 Datasets description and experimental setup

Different multi-view videos have been used in this investigation that includes Break-dancers, Ballet, Exit and Race1 datasets. These multi-view videos have been low pass filtered and decimated as mentioned earlier in subsections 4.1.2 and 4.2.2. The first seven views from each multi-view video are used, where monoscopic video is constructed through multiplexing frames that belong to these views using time-first ordering (Chen et al., 2009b).

4.3.3 Statistics of block matching among reference frames

Block matching is statistically analysed through determining on average how much each reference frame is used in predicting P-frame. Intra-prediction and all coding modes are enabled. Break-dancers dataset is compressed via H.264/AVC based multi-view video coding using target bitrate of 64 Kbps. Twenty-one reference frames have been used in predicting P-frame that belongs to middle view V_K as depicted in Figure 4-16-a, where T_0-T_2 and M_0-M_{14} are the temporal frames and spatiotemporal frames respectively, while S_0-S_2 are the spatial reference frames. The frames indices are sorted in opposite to coding order as depicted in Figure 4-16-b. The block matching statistics are computed during coding Break-dancers MVV. Table 4-7

shows this distribution among reference frames, e.g. block matching contribution for S_0 and M_9 are 66% and 0.37% respectively. From Table 4-7, the distribution of the block matching amongst reference frames is inconsistent with the position of the reference frames' indices in the buffer List 0. Hence, current bit allocation for representing each reference frame could be improved through sorting reference frames' indices in a suitable order.

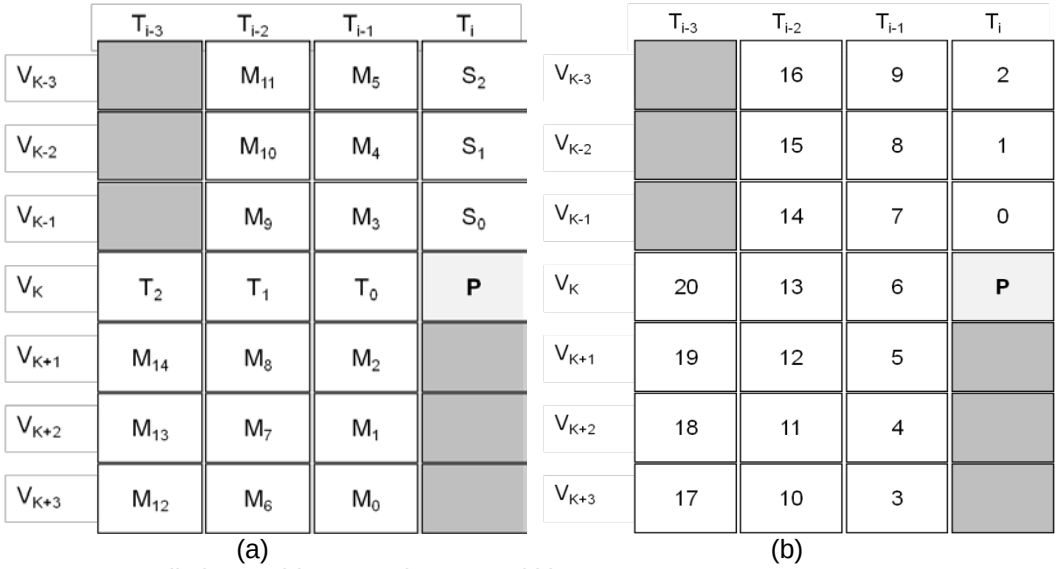


Figure 4-16 Prediction architecture a) RFS and b) RFO

Table 4-7 Statistics of block matching using opposite to coding order RFO (K=3)

View number	T_{i-3}	T_{i-2}	T_{i-1}	T_i
View 0		0.02	0.087	0.87
View 1		0.03	0.22	1.95
View 2		0.37	0.95	66
View 3	0.67	1.8	16.85	P
View 4	0.2	0.57	5.06	
View 5	0.08	0.26	1.67	
View 6	0.09	0.26	2.03	

In the second set of experiments, the reference frames indices are sorted according to their contributions of block matching beside their spatial position to the current frame as shown in Figure 4-17. Another statistical analysis of block matching is performed using the proposed reference frame indexing. The results for Break-dancers sequences are tabulated in Table 4-8. From this table, it can be seen that the temporal frames have higher role of block matching than spatial reference frames in addition to the majority of prediction came from recent temporal reference frame (T_0). This finding matches the fact that temporal correlations are higher than the

spatial correlations. The statistic of Skip and Intra-Prediction for using the first set of indexing reference frames are 40% and 7.6%, while for using the second indexing are 55.8% and 5.82%, respectively. The percentage of the macroblocks using the Skip mode prediction is increased by 15.8%. The encoded skipped macroblock cost a single bit to signal this mode instead of sending its prediction information. Additional coding performance would be achieved by using the proposed reference frame indexing. It can also be seen that the percentage of the Intra-coded macroblock is reduced by 1.78% which would improve coding performance since it is more costly than other coding modes.

	T_{i-3}	T_{i-2}	T_{i-1}	T_i
V_{K-3}		17	13	8
V_{K-2}		15	7	5
V_{K-1}		11	3	1
V_K	9	4	0	**P**
V_{K+1}	18	10	2	
V_{K+2}	19	14	6	
V_{K+3}	20	16	12	

Figure 4-17 RFO according to the reference frames contributions of block matching

Table 4-8 Statistics of block matching using the proposed RFO

View number	T_{i-3}	T_{i-2}	T_{i-1}	T_i
View 0		0.02	0.07	0.29
View 1		0.01	0.1	1.36
View 2		0.15	1.18	23.61
View 3	0.5	1.38	67.64	P
View 4	0.1	0.22	1.93	
View 5	0.08	0.17	0.73	
View 6	0.04	0.1	0.32	

4.3.4 Proposed prediction architectures

The outcome from the previous statistical analysis is used to derive the proposed prediction architecture. Figure 4-18 shows the proposed architectures using 4 and 6 reference frames. The frames indices are sorted in interleave order, where there is no preferable direction for sorting reference frames' indices (e.g. temporal and spatial directions). The numbers in each block represent reference frame index, where V_k,

V_{k+1}, V_{k-1} and V_{k-2} are current and its three corresponding neighbouring views. There is no typical prediction architecture for IPPP coding structure. Therefore, the proposed prediction architecture is evaluated alongside five different prediction architectures that use IPPP coding structure. The first two prediction architectures[23] (named Typical-A and, Typical-B) are based on the outcomes of *Merkle et al.* and *Kaup and Fecker* (Merkle et al., 2007a; Kaup & Fecker, 2006). The first Prediction Architecture (PA) gives higher priority to temporal reference frames while the second places the spatial reference frame first in List 0. The 3rd and 4th prediction architectures represent the prediction architectures (mode 1 and mode 3) proposed by *Sheikh Akbari et al.* that use similar RFS with different RFO (Sheikh Akbari et al., 2007). The 5th prediction architecture is proposed by *Fecker and Kaup* (Fecker & Kaup, 2005). These prediction architectures reflect three different reference frame selection categories. Typical-A and Typical-B relies in majority on temporal frames, while prediction architecture proposed by Fecker and Kaup relies on spatial frames. Prediction architectures proposed by *Sheikh Akbari et al.* use almost balance amounts for temporal and spatial frames. Break-dancers, Ballet, Exit and Race1 are coded using these prediction architectures.

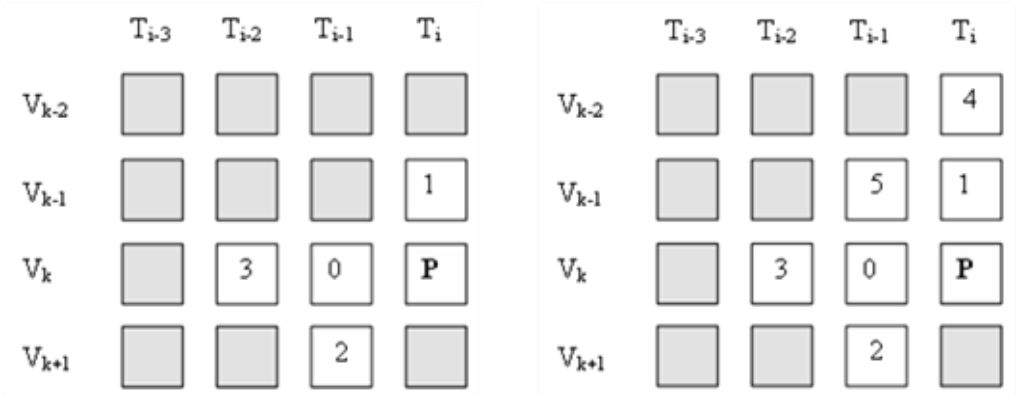


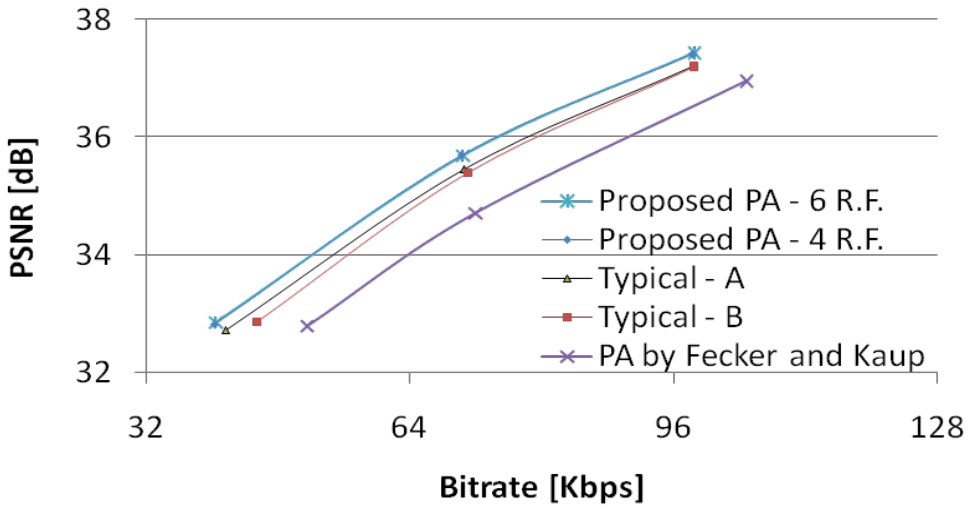
Figure 4-18 The proposed prediction architectures using: a) 4 reference frames and b) 6 reference frames

4.3.5 Results and discussions

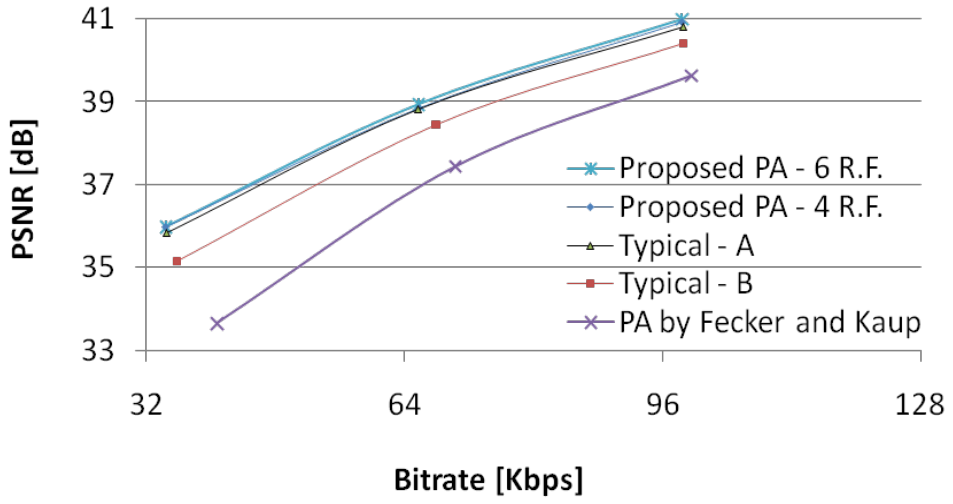
The Peak Signal to Noise Ratio (*PSNR*) measurement was used to assess the quality of the reconstructed luminance components of the decoded sequences. Figure 4-19 and Figure 4-20 show rate-distortion curves when coding Break-dancers, Ballet, Race1 and Exit datasets at low bitrates; 32, 64 and 96 kbps. From Figure 4-19, it can be seen that the proposed prediction architecture improves the

[23] They include the recent three temporal reference frames and nearest neighbouring spatial frame

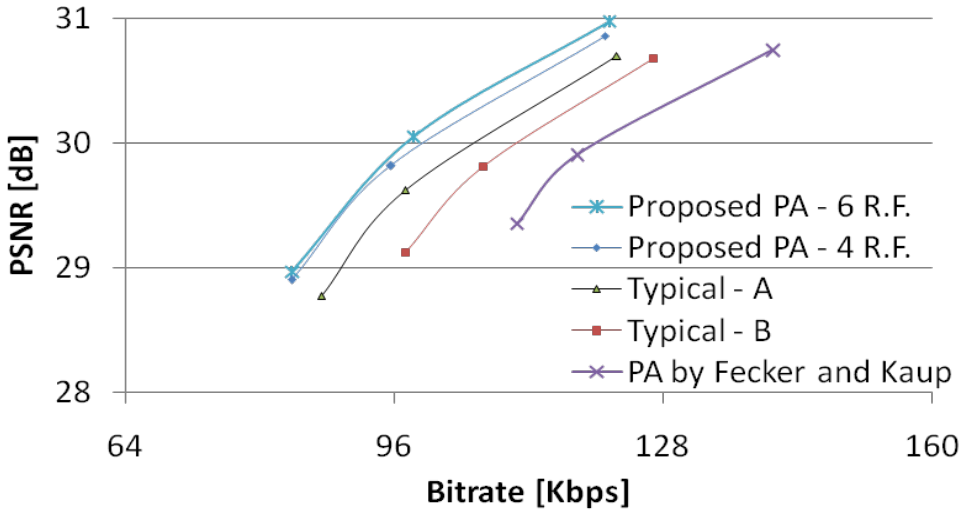
coding performance with respect to other prediction architectures that are based on IPPP coding structure. The codec using the proposed prediction architecture (six reference frames) provides significant coding gain up to 2.3 dB compared to the corresponding codec that deploys PA that is proposed by *Fecker and Kaup*. It improves coding gain using four reference frames by up to 0.43 dB and 0.83 dB compared to Typical-A and Typical-B prediction architectures respectively. From Figure 4-20, the proposed prediction architecture gives higher coding efficiency than the corresponding PA proposed by *Sheikh Akbari et al.* by up to 0.8 dB. From Figure 4-19, the proposed PA using 4 frames gets less coding gain improvement to Typical–A PA in comparison to the one presented by *Fecker and Kaup*. This indicates the importance of relying on temporal frames more than spatial frames (RFS) in addition to putting the index of the nearest temporal frame first in the List.



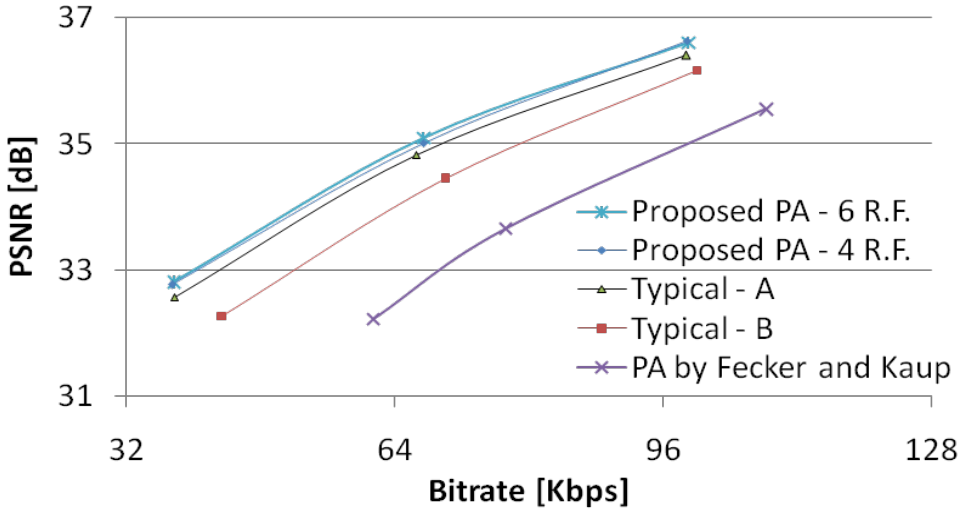
(a)



(b)

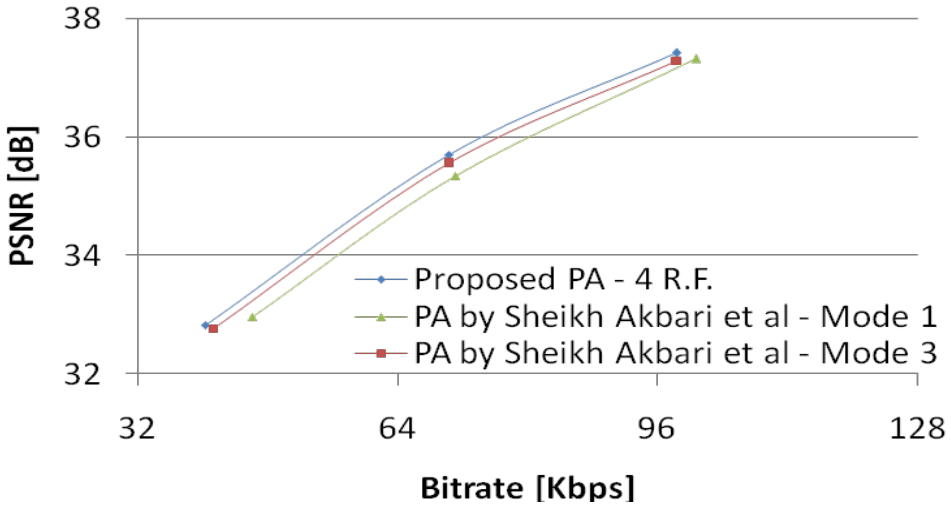


(c)

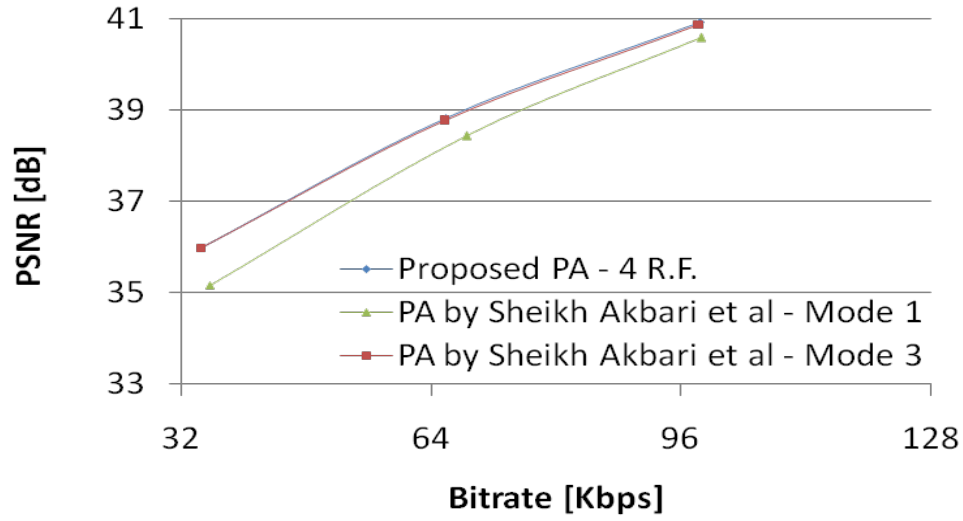


(d)

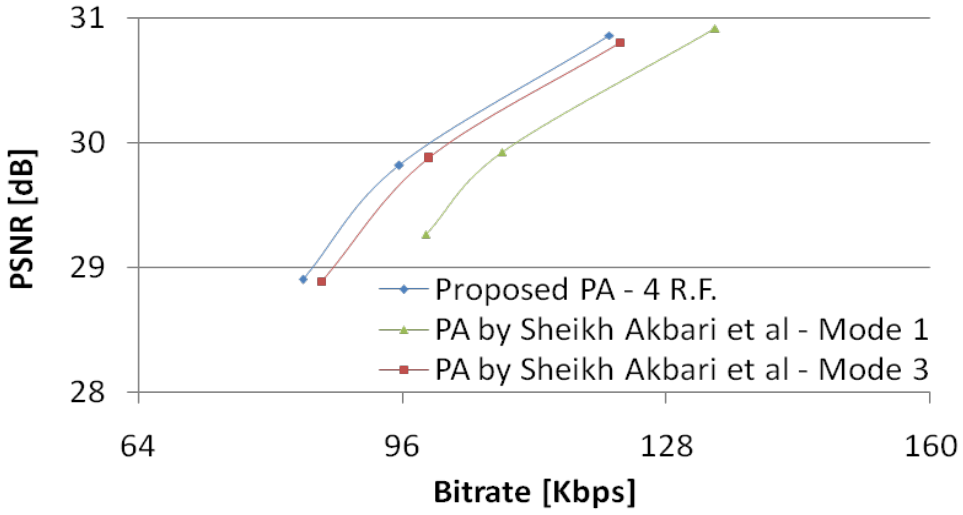
Figure 4-19 (a-d) Coding performance using proposed prediction architectures (4 and 6 reference frames) among three different prediction architectures for Break-dancers, Ballet, Exit and Race1 respectively



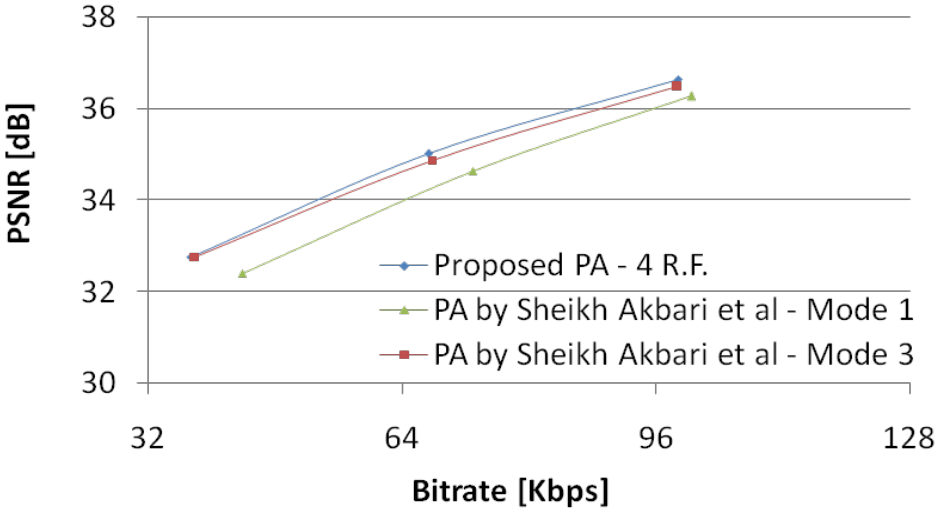
(a)



(b)



(c)



(d)

Figure 4-20 (a-d) Coding performance using the proposed PA (4 reference frames) and prediction architectures proposed by *Sheikh Akbari et al* for Break-dancers, Ballet, Exit and Race1 respectively

H.264/AVC relies on multi-reference frame and coding modes to provide flexibility during inter-picture prediction. Coding modes are analysed starting from block size of 16 × 16 to 4 × 4 for the proposed PA using six reference frames as shown in Figure 4-21. It is clear that the coding modes of sub-macroblock partitions are rarely used during coding Break-dancers at low bitrate. Since the majority of inter-picture prediction comes from few frames using different sizes of macroblock partition, a trade-off study among the reference frames and coding modes is applied. The next subsection presents the study in terms of computational complexity and coding performance of multi-view video coding.

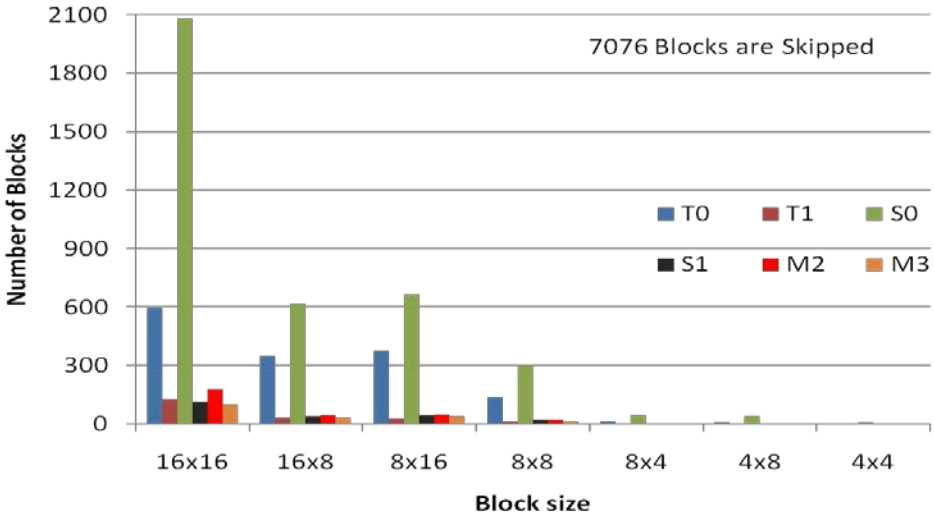


Figure 4-21 Mode distribution among reference frames at low bitrate

4.3.6 Computational complexity and coding performance trade-off study

Different sets of reference frames are selected and evaluated. Their coding modes are set into two categories based on mode distribution analysis, where the first category enables all coding modes while the second enables macroblock partition modes. There are six reference frames; they are T_0, T_1, S_0, S_1, M_1 and M_2. From these reference frames, three sets of reference frames are evaluated as shown in Figure 4-22, where white blocks are reference frame selection. These sets use the first two, four and six reference frames. They are named first, second and third reference frame selection; RFS-1, RFS-2 and RFS-3, respectively. The relation among them is: RFS-1 ⊆ RFS-2 ⊆ RFS-3. These sets provide 91.25%, 94.56% and 97.1% of block matching. If computational complexity for RFS-1 is *X*; then the corresponding complexity for 2nd and 3rd are roughly 2 and 3 multiplied by *X*.

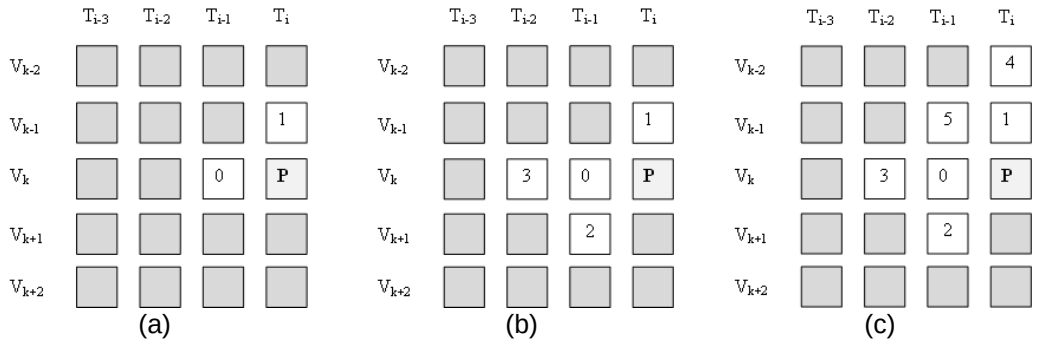


Figure 4-22 (a - c) show RFS-1 to RFS-3 respectively

Four Multi-view videos; Break-dancers, Ballet, Race1 and Exit, have been encoded twice via H.264/AVC based MVC at low bitrates. In the 1st phase, all coding modes are enabled while in the 2nd set of experiments; sub-macroblock partitions are disabled. Table 4-9 show the results for MVC using these reference frame selections when all coding modes are enabled. The computational complexity is realised by the average encoding time per frame. Table 4-10 shows the average encoding time among datasets using different set of reference frames. From this set of experiments, it can be observed that the average encoding time per frame is proportional with the number of reference frames. Δ*PSNR* when RFS-2 is used is in the range of -0.08 to -0.31 dB with respect to RFS-1. For six reference frames, the corresponding coding gain with respect to RFS-1 is in the range of -0.14 to -0.43 dB. Δ*PSNR* and Δ*BR* are defined by equation 4-4 and equation 4-5 respectively, where $PSNR_1$ and BR_1 represent coding performance when RFS-1 is used, while *i* is the label of RFS (2 and 3).

$$\Delta\, PSNR_i = PSNR_1 - PSNR_i, i \in 2\ and\ 3 \quad (4\text{-}4)$$

$$\Delta\, \mathrm{BR}_i = BR_1 - BR_i, i \in 2\ and\ 3 \quad (4\text{-}5)$$

Table 4-9 Coding performance when MVC uses all coding modes

MVV	**Break-dancers**		**Ballet**		**Race1**		**Exit**	
RFS	***PSNR***	**Bitrate**	***PSNR***	**Bitrate**	***PSNR***	**Bitrate**	***PSNR***	**Bitrate**
1	34.55	70.1	38.34	65.94	29.54	94.06	34.53	66.18
2	34.63	70.03	38.49	66.37	29.85	95.23	34.70	67.25
3	34.69	70.23	38.61	66.02	29.97	95.42	34.72	67.97

Table 4-10 Average encoding time (seconds) per frame using all coding modes

RFS	**Break-dancers**	**Ballet**	**Race1**	**Exit**
1	54	53	37	39
2	111	104	83	77
3	161	152	101	108

Table 4-11 shows the results when MVC uses macroblock partitions, while the corresponding average encoding time per frame is shown in Table 4-12. In the second set of experiments, disabling sub-macroblock partition coding modes speeds up MVC by on average 26% of the encoding time, while the $\Delta PSNR$ is dropped by on average 0.1 dB. Coding multi-view video using six reference frames through enabling all coding modes gets higher coding gain (by up to 0.34 dB) than the corresponding RFS that relies on the nearest temporal and spatial frames alongside macroblock partitions. This is achieved at expense of increasing the required encoding time to compress the given MVV, where RFS-3 needs on average 3.8 times more the required time taken by the RFS-1.

Table 4-11 Coding performance using macroblock partition sizes

MVV	**Break-dancers**		**Ballet**		**Race1**		**Exit**	
RFS	***PSNR***	**Bitrate**	***PSNR***	**Bitrate**	***PSNR***	**Bitrate**	***PSNR***	**Bitrate**
1	34.41	68.09	38.243	65.65	29.536	94.79	34.423	65.72
2	34.655	70.12	38.332	65.75	29.729	94.9	34.546	65.82
3	34.664	70.06	38.392	65.83	29.898	95.3	34.596	65.91

Table 4-12 Average encoding time (sec) per frame using macroblock partition sizes

RFS	**Break-dancers**	**Ballet**	**Race1**	**Exit**
1	39	39	28	28
2	79	77	59	58
3	114	112	83	80

4.3.7 Conclusions

Block matching statistical analysis is used to reveal reference frames that have significant contribution of block matching for MVC, where prediction architectures using four and six reference frames are proposed. Reference frame selection using six frames includes nearest two temporal, two spatial and two spatiotemporal frames while interleaved order is used to sort their reference frame indices. The proposed architecture outperforms a set of prediction architectures that use IPPP coding structure by coding gain up to 2.3 dB. The application of the prediction architecture with smaller number of reference frames is preferred at low bitrate. Trade-off study among coding efficiency and computational complexity using different set of reference frame selections and coding modes is evaluated. For low computational complexity MVC, nearest temporal and spatial frames are deployed in RFS while coding modes would include macroblock partitions.

This section used static reference frame ordering during coding MVV. The next section will investigate efficient mechanism for reference frame reordering.

4.4 Adaptive reference frame ordering algorithm

4.4.1 Introduction

Multi-view videos have different degree of scene complexity and motion among existing objects in the scene. Therefore, it is difficult to unify reference frame selection or reference frame ordering for all multi-view videos. E.g. temporal frames can be more dominant than other frames as in Ballet dataset or almost balanced between temporal and spatial frames as in Break-dancers dataset.

Reordering reference frames indices inside List 0 buffer would reduce the total number of bits required to signal reference frames indices by assigning shorter indices for most frequent reference frames. This saves unnecessary bits when signalling these indices that consequently lead to improve the coding performance of MVC. Therefore, this section focuses on developing efficient mechanism for reference frame reordering. The following subsection will exploit observation from reference frame ordering using Microsoft datasets. Based on this observation, adaptive reference frame ordering algorithm will be proposed. Different applications are then discussed, where significant remarks will be concluded.

4.4.2 Multi-view video coding using static reference frame order

This subsection investigates how to predict suitable reference frame ordering (RFO) at each time slice. A statistical analysis of block matching among reference frames has been conducted using prediction architecture depicted in Figure 4-23, where the reference frame ordering is static. This analysis determines the contribution of each reference frame for predicting P-frame using all coding modes. This analysis is deployed during encoding Break-dancers and Ballet at low bitrate.

The basic idea behind this subsection is to reveal the order of reference frames after encoding the P-frame using the following order; T_0, T_1, S_0 and S_1. The statistic of the block matching amongst reference frames is calculated and used to sort the reference frames in descending order. The sorted reference frames are then given a label. These labels are tabulated in Table 4-13. There are six reference frame orders starting from Label A to Label F and Ref_i represent either temporal (T) or Spatial reference frame (S). The first seven views are used during coding each multi-view video, where the first, two views; V_0 and V_1; are not used in this analysis due to

unavailability of some reference frames (e.g. S_0 and S_1). Table 4-14 and Table 4-15, show the suitable reference frames order in terms of "labels", based on the statistics of block matching among four reference frames for the first 55 frames from time slice t_2 to t_{12}. E.g. RFO labels for the frame that belongs to t_2 for Break-dancers and Ballet are 'C' and 'A' respectively. It is worth mentioning that reference frames orders labelled by 'A' and, 'B' are similar because their first two reference frames are the same (T_0 then S_0) and they always have the most contribution of block matching prediction (the same concept applies to labels 'C' and 'D'). The shaded cells in both tables show consecutive frames within the same view (temporal frames) which should be coded using different reference frame orders. Also, it can be inferred that the suitable reference frame order would be predicted in most cases, using the previous temporal frame.

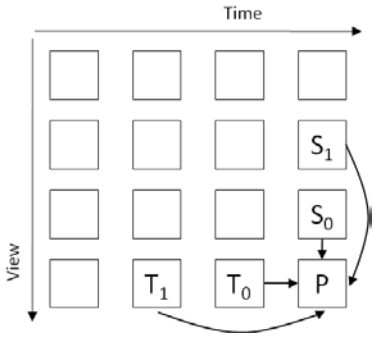


Figure 4-23 Prediction architecture used in investigating reference frame order

Table 4-13 Reference frame orders tagged with different labels

Case	Ref_0	Ref_1	Ref_2	Ref_3
A	T_0	S_0	S_1	T_1
B	T_0	S_0	T_1	S_1
C	S_0	T_0	S_1	T_1
D	S_0	T_0	T_1	S_1
E	T_0	T_1	S_0	S_1
F	S_0	S_1	T_0	T_1

Table 4-14 Labels that reflect the suitable RFO for Break-dancers

V_k	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}
V_2	C	C	C	C	D	C	D	E	D	C	C
V_3	B	B	B	B	B	B	A	B	B	B	B
V_4	C	D	C	C	C	C	C	C	C	C	C
V_5	A	A	A	C	C	C	C	C	C	C	C
V_6	C	C	C	C	C	C	C	C	C	C	C

Table 4-15 Labels that reflect the suitable RFO for Ballet

V_k	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}
V_2	A	C	D	B	B	B	B	B	B	B	A
V_3	B	B	B	A	B	A	A	A	A	A	B
V_4	A	B	B	B	A	A	B	A	A	B	A
V_5	B	A	B	A	D	B	D	D	C	B	C
V_6	B	A	B	C	C	C	C	C	C	C	C

4.4.3 Proposed adaptive reference frame ordering algorithm

The previous subsection shows that the RFO for current P-frame is mostly predicted through a recent temporal frame. Therefore, adaptive reference frame ordering algorithm is proposed that is depicted in Figure 4-24. For a P-frame, it checks first if the frame is located in a position, where partial reference frames are available (transient state e.g. all P-frames in the first time slice; t_0). In this stage, the algorithm uses predefined prediction architecture to encode the frame.

In a non-transient scenario, the algorithm loads the corresponding order of reference frames then encodes the P-frame using that order. After that, the algorithm loops on all its macroblocks to compute the block matching statistics among all reference frames. When there is no scene change, the algorithm sorts the reference frames based on their block matching statistics and its new order will be stored and applied to the next temporal frame.

When the video codec compresses frames that belong to a new scene, the majority of macroblocks that belong to the first frame in the new scene are intra-predicted. Hence the algorithm relies on the amount of intra-coded macroblocks to detect scene changes. If the percentage of intra-predicted macroblocks exceeds certain threshold (60%), then the following P-frames will use similar reference frames order to the corresponding P-frames in the transient state (Brandt et al., 2008). Therefore, the next frame to be coded that is located within the same time slice will use short indices for spatial reference frames through placing the order of these frames first in List 0.

The algorithm would be deployed in both, encoder and decoder where the encoder does not need to signal the new reference frame ordering when it occurs. The decoder computes block matching statistics using current reference frame ordering during motion and disparity compensation. After decoding the current frame, the reference frames' indices are sorted according to the block matching statistics. The decoder stores new reference frame ordering to be used for the next temporal frame as long as there is no scene changes among decoded frames.

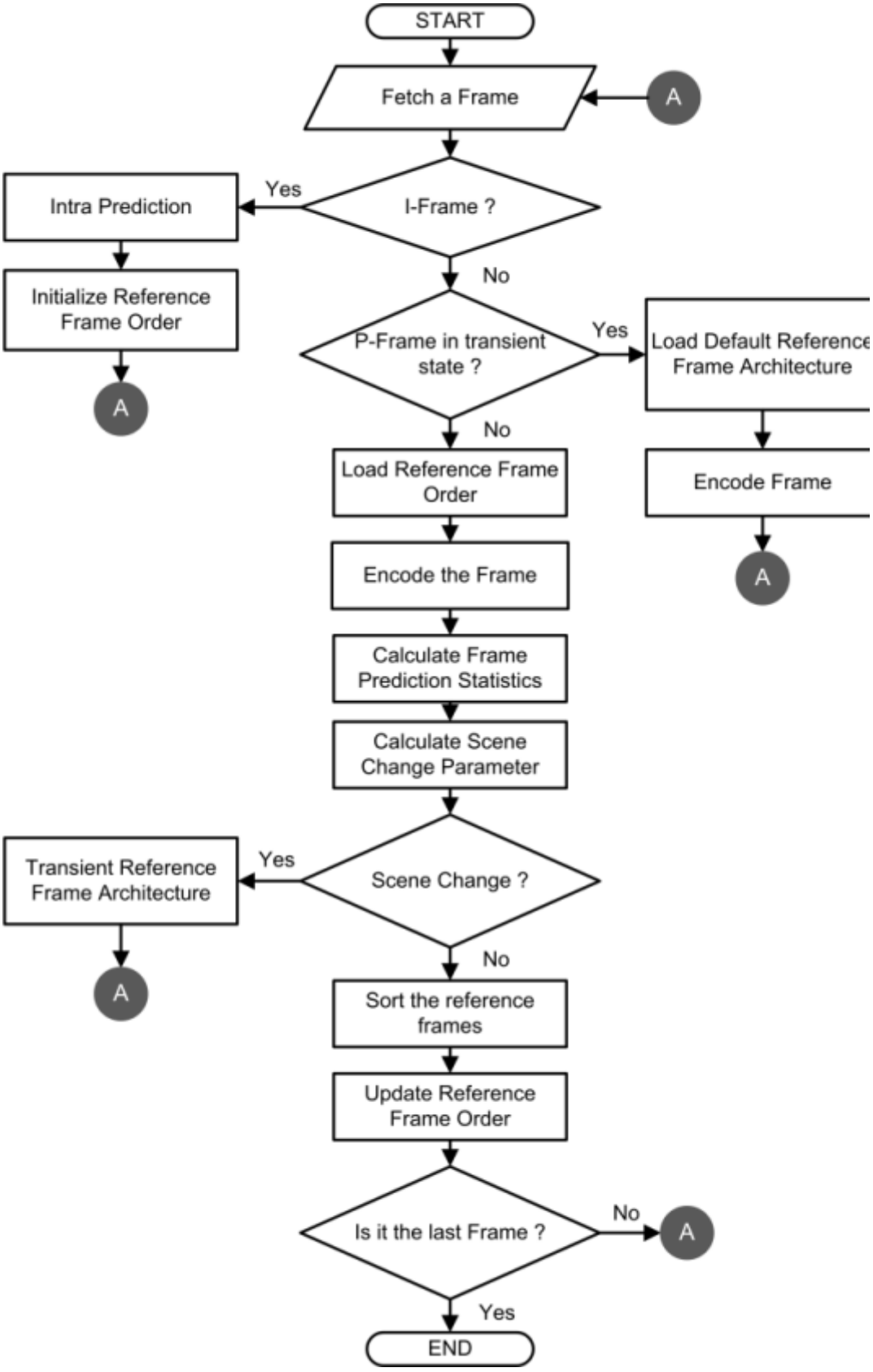


Figure 4-24 Adaptive reference frame ordering algorithm

4.4.4 Proposed algorithm applications

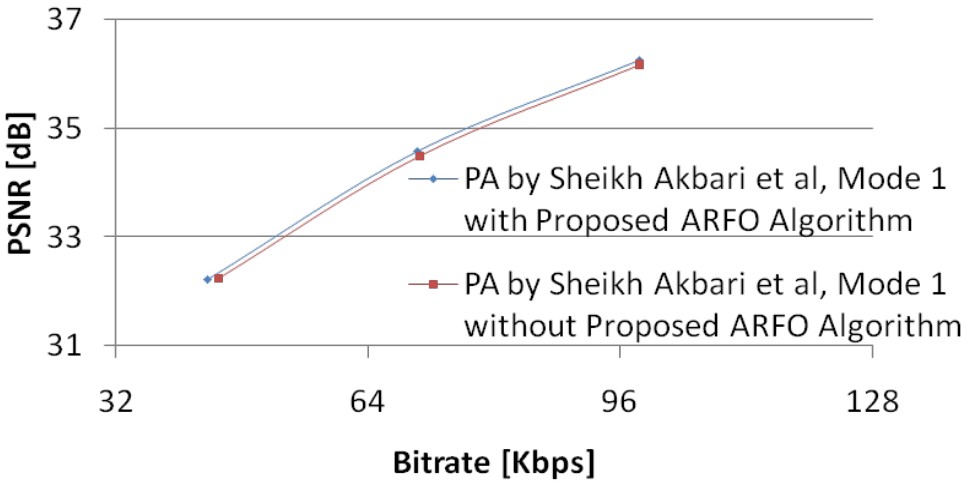
There are two applications that would benefit from the proposed Adaptive Reference Frame Ordering algorithm (ARFO). The first application is coding given multi-view video by PA that contains many reference frames, while the latter is coding a sequence that contains hard scene changes.

The proposed algorithm is applied to prediction architectures (mode 1 and mode 3) proposed by *Sheikh Akbari et al.* to code four multi-view videos, Break-dancers, Ballet, Race1 and Exit (Sheikh Akbari et al., 2007). These architectures use five reference frames, where their RFO are clearly stated. These architectures use three temporal, spatial and spatiotemporal reference frames. The first architecture (mode 1), places spatial and spatiotemporal frames indices first, while the second architecture (mode 3), places temporal frames in the beginning of the other reference frames. The proposed algorithm starts with the same RFO that is defined in each mode. Frames located in the time slice below t_3 will be coded using the available reference frames (transient state). After t_3, the algorithm starts to adapt the reference frame ordering dynamically.

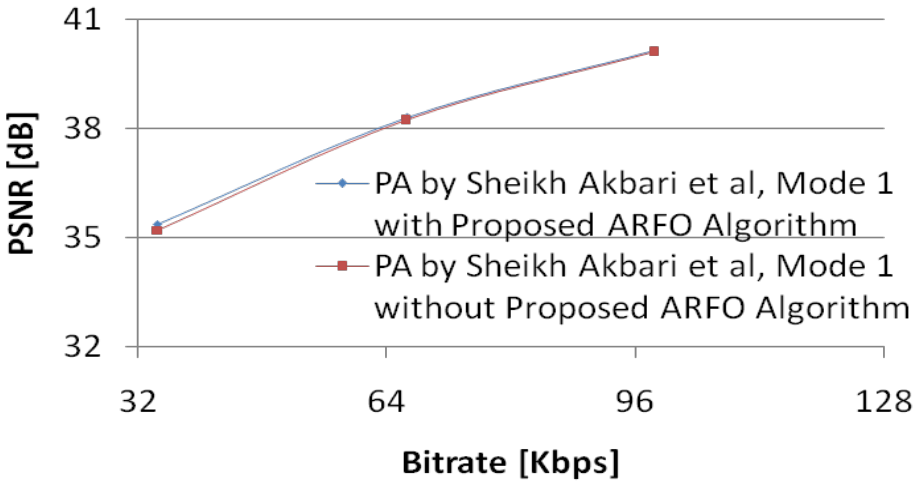
The proposed algorithm is validated also by integrating it to PA; mode 3 proposed by *Bilen et al.* when coding a sequence that contains hard scene changes (Bilen et al., 2006). This architecture defines clearly RFO and it uses a smaller number of reference frames than prediction architectures proposed by *Sheikh Akbari et al*. The sequence is generated from Break-dancers, Ballet, Race1 and Exit multi-view videos. To generate sequence with hard scene changes, Microsoft datasets are decimated first to QVGA to match the spatial-resolution of both KDDI and MERL datasets. Then, the first six frames from each view within each dataset are used to generate the sequence where sixteen consecutive frames from each video are concatenated to form multi-view video sequence, thus the resulted sequence contains 192 frames.

4.4.5 Results and discussions

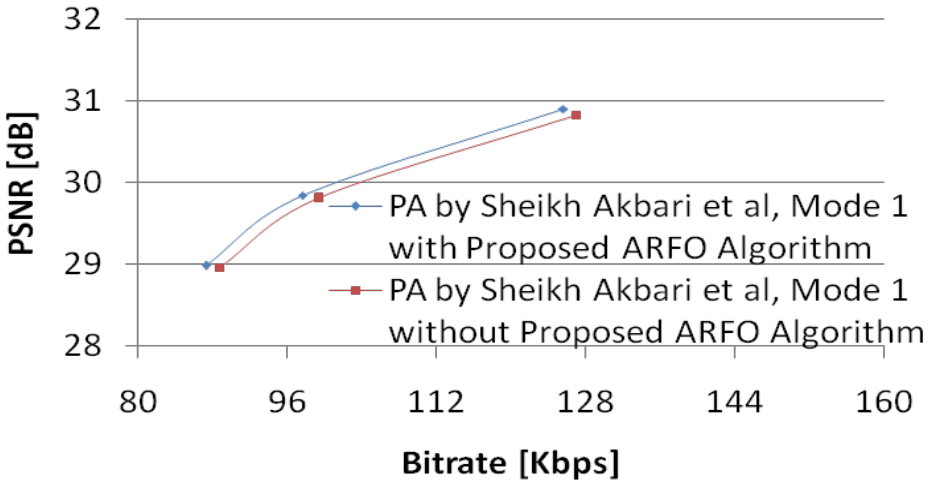
Figures 4-25 and 4-26 show the coding performance of the multi-view video codec using the proposed adaptive reference frame ordering algorithm in comparison to static reference frame ordering proposed by *Sheikh Akbari et al*. It can be seen that the proposed algorithm gives higher coding performance compared to the static reference frame ordering by up to 0.2 dB. From these figures, the ARFO algorithm is less efficient when applying to videos with dominant temporal correlation (e.g. Ballet and Exit videos).


37
35
33
31
PSNR [dB]
32
64
96
128
Bitrate [Kbps]
PA by Sheikh Akbari et al, Mode 1 with Proposed ARFO Algorithm
PA by Sheikh Akbari et al, Mode 1 without Proposed ARFO Algorithm

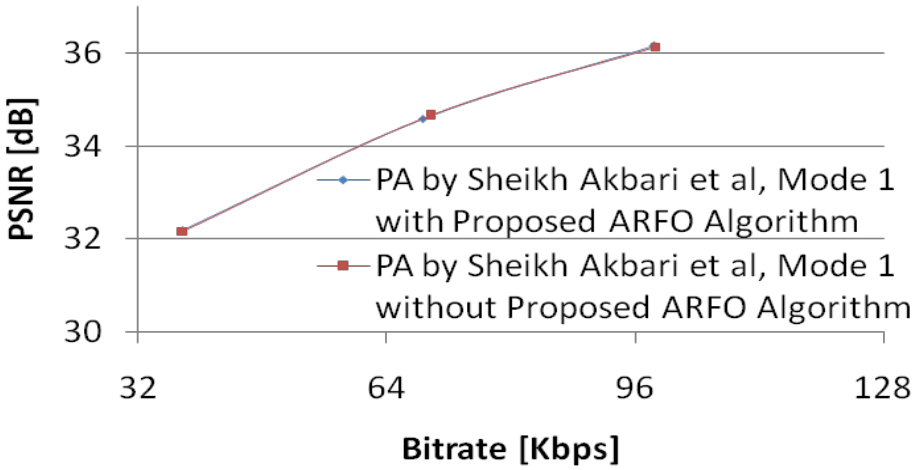
(a)


41
38
35
32
PSNR [dB]
32
64
96
128
Bitrate [Kbps]
PA by Sheikh Akbari et al, Mode 1 with Proposed ARFO Algorithm
PA by Sheikh Akbari et al, Mode 1 without Proposed ARFO Algorithm

(b)

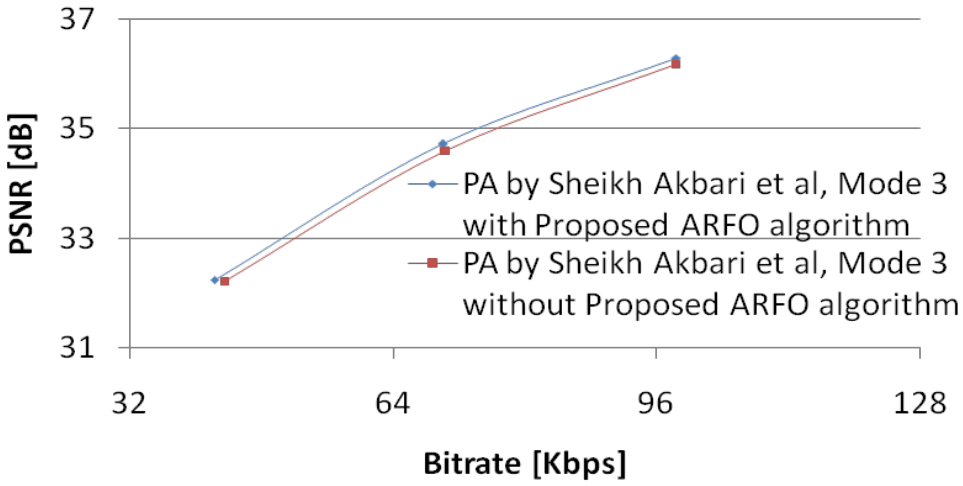

32
31
30
29
28
PSNR [dB]
80
96
112
128
144
160
Bitrate [Kbps]
PA by Sheikh Akbari et al, Mode 1 with Proposed ARFO Algorithm
PA by Sheikh Akbari et al, Mode 1 without Proposed ARFO Algorithm

(c)

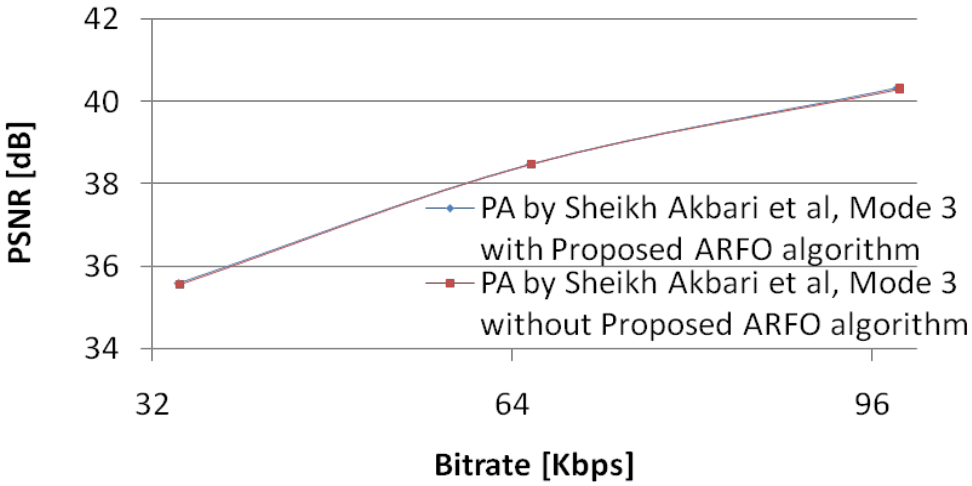


(d)

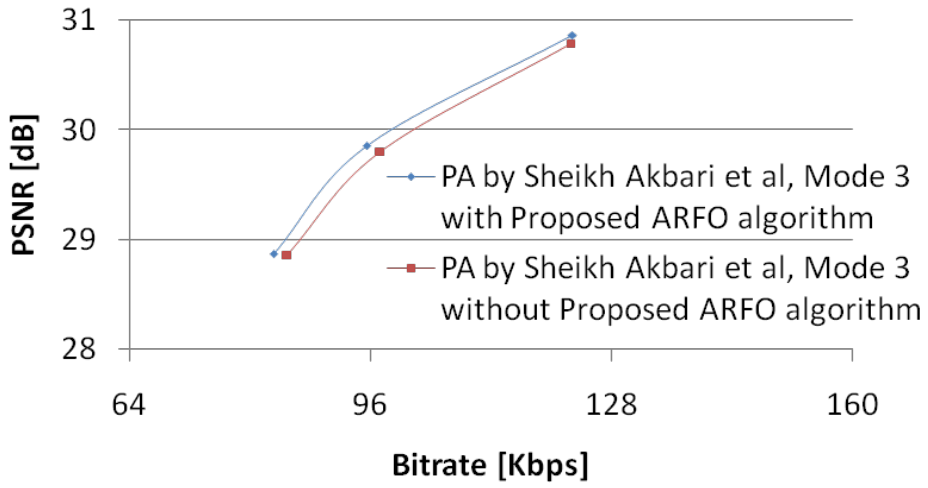
Figure 4-25 (a-d) Coding performance using the proposed algorithm when the PA (mode 1) proposed by *Sheikh Akbari et al.* is used for Break-dancers, Ballet, Exit and Race1 respectively (Sheikh Akbari et al., 2007)



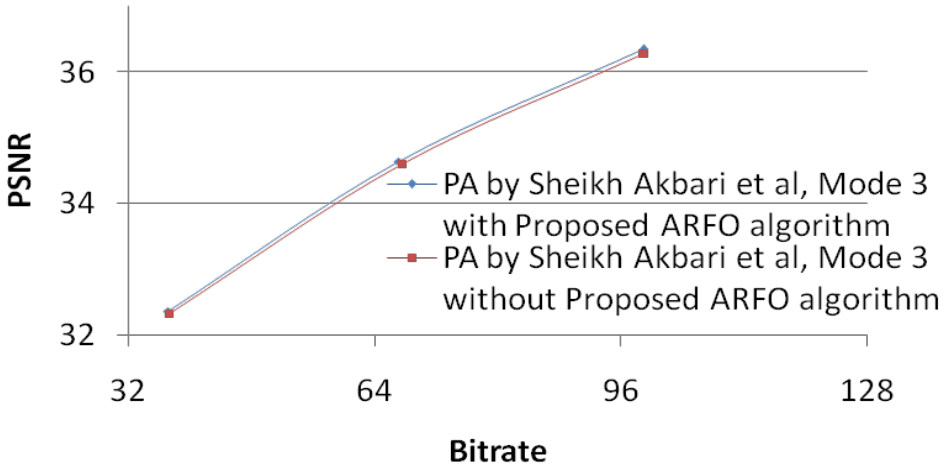
(a)



(b)



(c)



(d)

Figure 4-26 (a-d) Coding performance using the proposed algorithm when the PA (mode 3) proposed by *Sheikh Akbari et al.* is used for Break-dancers, Ballet, Exit and Race1 respectively (Sheikh Akbari et al., 2007)

In the second application, the proposed algorithm is evaluated when coding sequence which contains hard scene changes. Results are shown in Figure 4-27 and Figure 4-28, where BR, BA, EX and, RA stand for Break-dancers, Ballet, Exit and Race1 respectively. From these figures, the proposed algorithm improves coding performance compared to the use of static reference frame ordering when scene change occurs. The proposed algorithm saves significant bitrates, up to 6.2% with respect to static reference frame ordering.

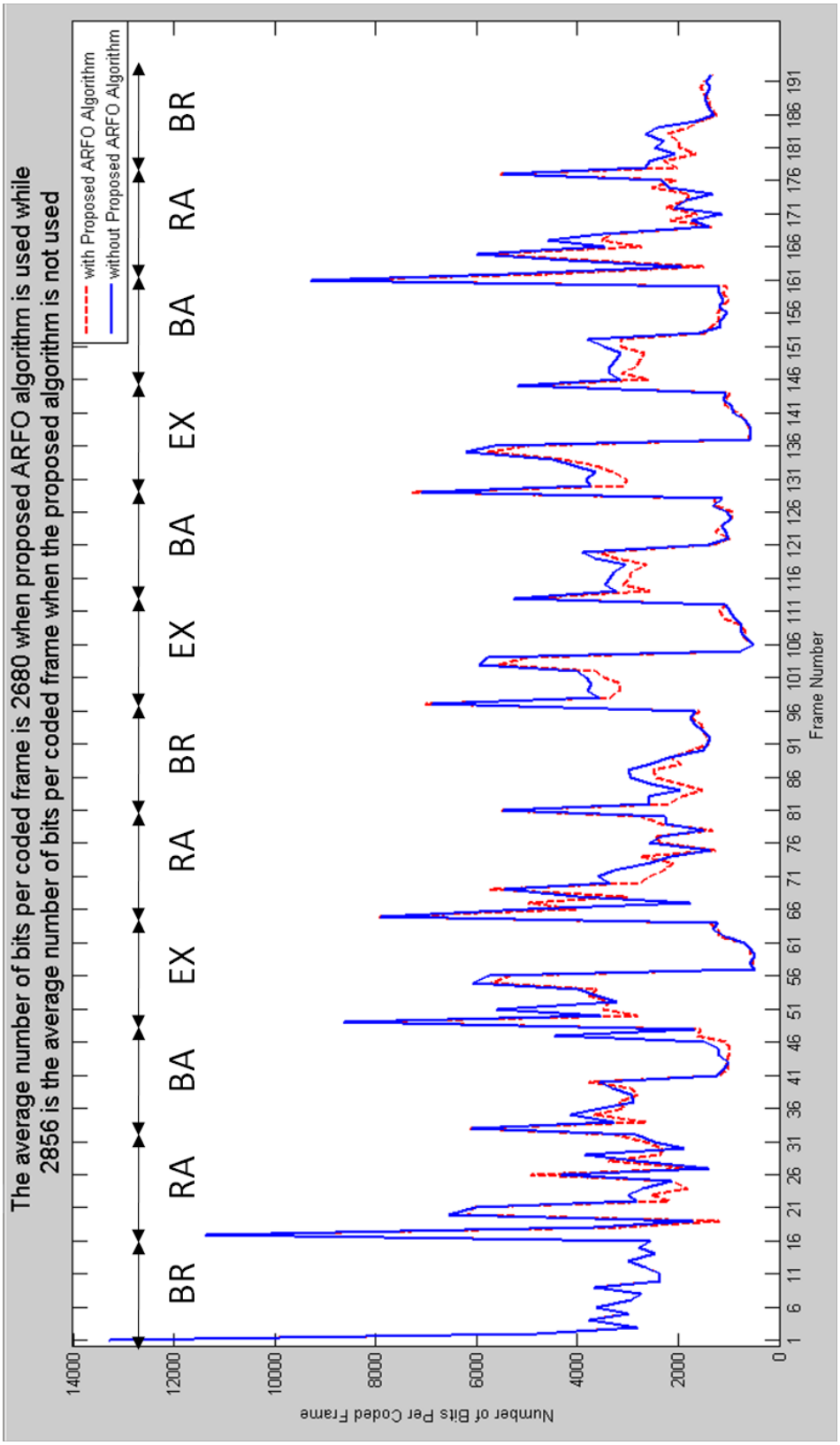


Figure 4-27 Number of bits per coded picture when ARFO algorithm is used with prediction architecture proposed by *Bilen et al.* (Bilen et al., 2006)

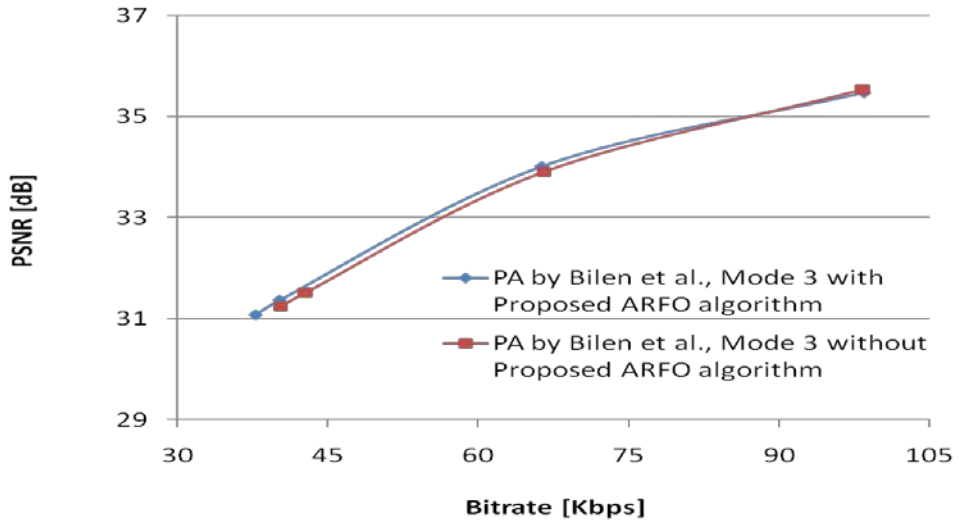


Figure 4-28 Coding performance using the proposed ARFO algorithm when the prediction architecture proposed by *Bilen et al.* is used (Bilen et al., 2006)

4.4.6 Conclusions

In this section, an adaptive reference frame ordering algorithm is proposed. It updates reference frames ordering through placing reference frame indices that have significant role of block matching first inside List 0. Reference frame order is predicted by analysing block matching statistics for previous temporal frame, where suitable order is exploited. This saves un-necessary bits required for addressing reference frame indices. Therefore the algorithm improves the coding performance for the prediction architecture which relies on multiple reference frames (up to 0.2 dB). When a video contains hard scene changes, the proposed algorithm updates reference frame order through placing spatial reference frames first. Hence the proposed algorithm saves significant amount of bits up to 6.2%.

The outcomes for the investigations applied for symmetric spatial-resolution multi-view video coding are summarised in the next section.

4.5 Summary of the investigations

The coding performance for multi-view video codec depends on block matching efficiency that exploits inter-view correlations among neighbouring views. Impact of camera separation on the coding performance of multi-view video coding is investigated. Based on coding results for wide baseline convergent multi-view video, inter-camera angle does not provide sufficient criterion to be used for selecting a suitable coding solution for a given multi-view video. Scene complexity has major

effect on inter-camera angle threshold. The dataset with a dominant temporal correlation has lower threshold (4°) than the dataset with balanced spatial-temporal correlation (12°).

Prediction architectures are investigated mainly to identify reference frame selection when H.264/AVC based MVC operates at low bitrates. Statistical analysis of block matching among reference frames is used to derive reference frame selection and their reference frame ordering. Based on block matching statistics, prediction architectures have been proposed, using four and six reference frames. The reference frame selection that uses six frames includes the nearest two temporal, two spatial and two spatiotemporal frames. Interleaved RFO is used to sort reference frames indices. The proposed prediction architectures yield superior coding performance than other prediction architectures, by coding gain up to 2.3 dB. Since few reference frames with a subset of coding modes have the majority of block matching contributions, a trade-off study among coding efficiency and computational complexity is conducted. For low computational complexity MVC, the nearest temporal and spatial frames to current P-frame are used for RFS while coding mode deploys only macroblock partitions.

Reference frame reordering is studied to provide an effective solution for ordering indices of reference frames. Based on block matching statistic results, the RFO would be predicted usually through the previous temporal frame. When the scene changes, reference frames indices are reordered in a way that places spatial reference frames first in List 0. An adaptive reference frame ordering algorithm is proposed. The proposed algorithm is tested among other prediction architectures that use static RFO. When coding MVV using multiple reference frames, it gets a coding gain up to 0.2 dB with respect to prediction architectures that are proposed by *Sheikh Akbari et al*. It saves bitrate up to 6.2% with respect to the PA that is proposed by *Bilen et al*. when coding MVV that has multiple scene changes.

Mixed spatial-resolution MVC is investigated in the next chapter. This coding approach is an efficient solution when coding MVV at low bitrates. The next chapter will focus on studying the effect of inter-view prediction direction on the coding performance of MVC. Different decimation and interpolation methods will be evaluated in terms of coding gain and computational complexity. Prediction architectures will be investigated for reference frame selection and reference frame ordering through analysing block matching statistics among the reference frames. The feasibility of improving visual quality for the coded low spatial-resolution frames will be explored in order to reduce blurriness artefacts at the receiver side.

CHAPTER 5. MIXED SPATIAL-RESOLUTION MULTI-VIEW VIDEO CODING

This chapter provides a set of studies toward investigating prediction architectures for asymmetric (mixed) spatial-resolution multi-view video coding when it operates at low bitrates. Particularly, the inter-view prediction is the main point for the investigations, where frames that belong to neighbouring views have different spatial-resolution (two interleaved sets of views that have frames with different spatial-resolution). Figure 5-1 shows the block diagram for the studies reported in this chapter. Circles labelled 1, 2, 3 and 4 reflect studies that investigate impact of inter-view prediction direction, explore different decimation and interpolation methods, derive RFS and RFO and investigate how to enhance visual quality for low spatial-resolution frames.

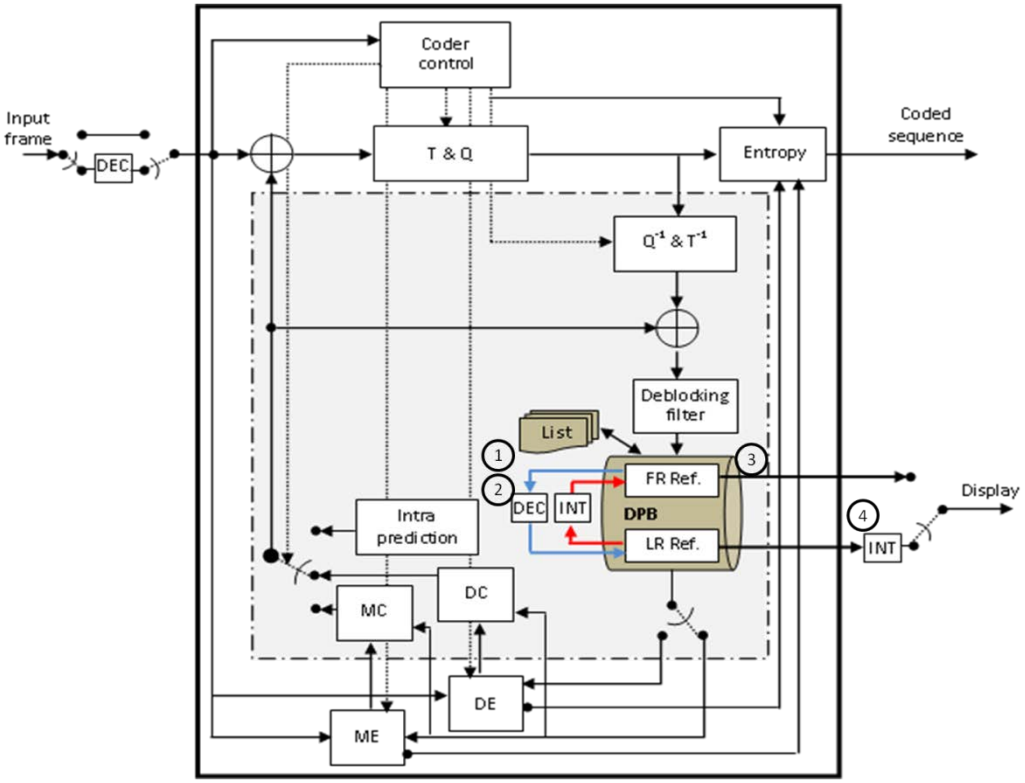


Figure 5-1 Block diagram for the studies conducted in mixed spatial-resolution multi-view video coding

Figure 5-2 shows the pre-processing and post-processing needed for mixed spatial-resolution multi-view video coding. At the sender side, frames that belong to subset of views are decimated by filtering and down-sampling prior to coding. At the

receiver side, the sequence is decoded, where low spatial-resolution frames are interpolated by up-sampling and filtering.

The first investigation explores the effect of using a low spatial-resolution frame to predict its neighbouring full spatial-resolution frame and vice versa. This would provide clear insight about impact of resolution reduction on inter-view prediction. Since two views are sufficient to conduct this investigation, stereoscopic video coding is used. Different methods in decimation and interpolation reference frames are evaluated in terms of coding gain and computational complexity. Statistical analysis of block matching will be conducted, where the results revealed by the corresponding analysis for symmetric multi-view video coding are used to identify reference frames candidates. Based on this statistical analysis, both RFS and RFO are derived. Multi-view videos have different characteristics of disparity, objects' motion and texture complexity. Therefore another detailed statistical analysis is applied to derive the correlation among neighbouring views; that entails omitting reference frames from RFS that would have insignificant amount of block matching. Adaptive reference frame ordering algorithm (reported in section 4.4), is deployed for mixed spatial-resolution multi-view video coding in order to evaluate coding efficiency when coding sequence that contains hard scene changes. The last section looked into enhancing the visual quality for coded low spatial-resolution frames. By exploiting information that exists in the neighbouring full spatial-resolution frame, the amount of blurriness could be minimised.

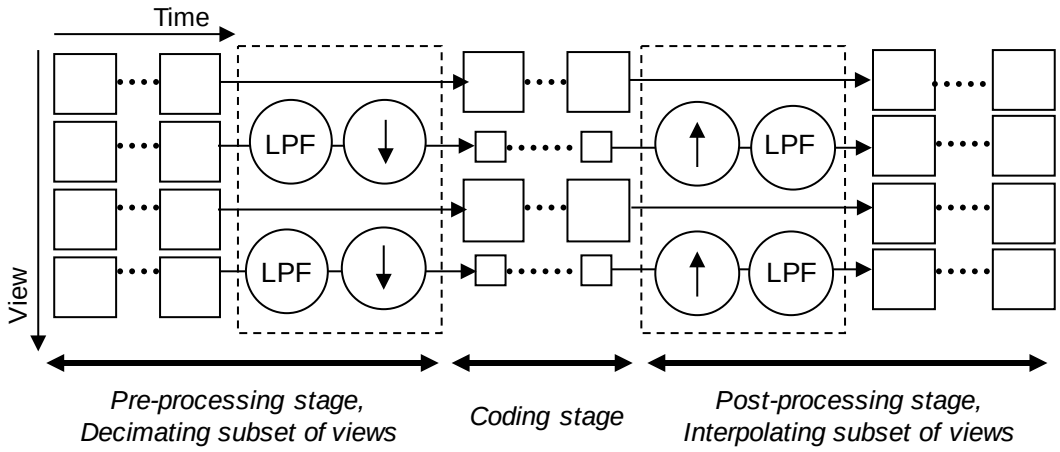


Figure 5-2 Mixed spatial-resolution multi-view video structure

The following section investigates the effect of inter-view prediction direction on the coding performance of stereoscopic video coding.

5.1 Impact of inter-view prediction direction on the coding performance of stereoscopic video coding

5.1.1 Introduction

Mixed spatial-resolution MVC implies additional process to be applied for reference frames prior to Inter-View Prediction (IVP), where frames' spatial-resolutions have to be similar before applying block matching. Figure 5-3 demonstrates different cases of IVP among reference frames with different spatial-resolution, where FR and LR stand for Full spatial-Resolution and Low spatial-Resolution frames. In this Figure, shaded block with dark grey represents reference frame while shaded block with light grey represents current frame to be coded via IVP. The first two cases are IVP, where REference Frame (REF) has to be decimated or interpolated in order to be used as a source to predict low spatial-resolution or full spatial-resolution frame respectively. The 3rd case is IVP with symmetric spatial-resolution (REF is used directly in block matching). Figure 5-4 shows IVP among asymmetric spatial-resolution frames. Figure 5-4-a illustrates the process of predicting low spatial-resolution frame, where filtering[24] full spatial-resolution REF and down-sampling are applied prior to block matching. Figure 5-4-b shows the process needed to predict FR frame via LR frame that involves up-sampling and filtering[25].

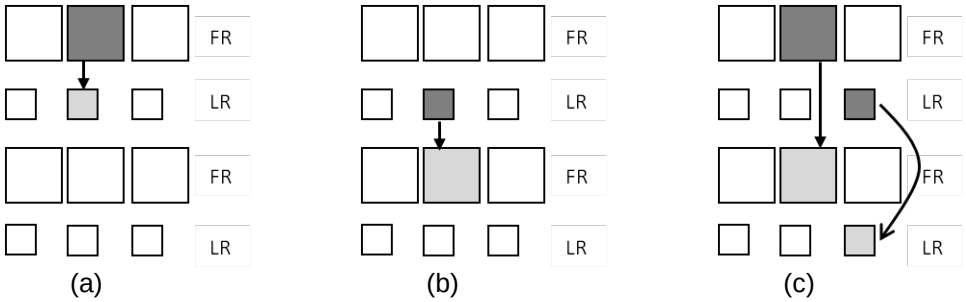
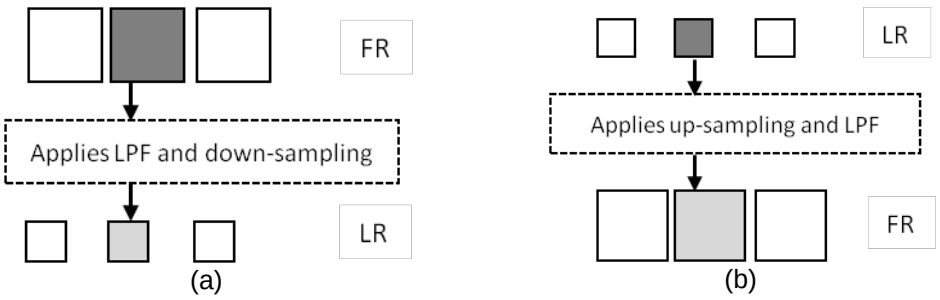


Figure 5-3 Different inter-view prediction for mixed spatial-resolution MVC



[24] It uses the same filter that is applied in decimating un-coded frames in the pre-processing stage
[25] It uses the same filter that is applied in interpolating coded frames in the post-processing stage

Figure 5-4 Two cases where reference frames have to be decimated or interpolated

In the following subsections, the effect of prediction direction is studied among mixed spatial-resolution stereoscopic video coding, where symmetric quality is applied when quantising transformed residual coefficients. The following subsection describes the datasets and pre-processing steps.

5.1.2 Mixed spatial-resolution stereoscopic videos preparation

Six multi-view videos have been used in this chapter. They are Break-dancers, Exit, Race1, Akko & Kayo, Ballroom and Rena. These videos are available online and have been recommended in the common test conditions of multi-view video coding (Su et al., 2006). Break-dancers, Exit and Race1 have been described earlier in subsections 4.1.2 and 4.2.2. Table 5-1 describes the other videos in terms of camera setup and frame rate. These videos are provided as YUV 4:2:0, where frame spatial-resolution is VGA.

Table 5-1 Datasets description

Sequence	Number of Cameras	Camera Setup	Camera Separation (cm)	Frame rate	Provider
Ballroom[26]	8	1D linear	19.5	25	MERL
Akko & Kayo	100 (5 x 20)	2D array	5	30	Nagoya university
Rena[27]	8	1D linear	5	30	Nagoya university

Mixed spatial-resolution stereoscopic videos have been pre-processed through two stages. Two views are selected from each MVV, where their spatial-resolution is first decimated by a factor of two[28] in the horizontal and vertical directions. This would provide sequences that fit requirements for low bitrate applications. The original spatial-resolution of the luminance components have been down-sampled using MPEG-4 down-sampling filter for all YUV videos. Break-dancers sequence has been converted from RGB into YUV format prior to down-sampling all colour channels. The choice of down-sampling filter was recommended in several studies (Chen et al., 2008a; Brust et al., 2010; Smirnov et al., 2010b). The outputs from the decimation are considered as full spatial-resolution videos. The second stage focuses on generating mixed spatial-resolution stereoscopic videos, where the target view that

[26] Online: ftp://ftp.merl.com/pub/avetro/mvc-testseq/orig-yuv/ballroom/
[27] Akko & Kayo and Rena datasets are available online: http://www.tanimoto.nuee.nagoya-u.ac.jp/
[28] It has been subjectively shown that asymmetric stereoscopic video coding using factor of two provides similar perceived quality to symmetric stereoscopic video coding (Aflaki et al., 2010)

should have low spatial-resolution frames is obtained through applying additional decimation by a factor of two in both spatial directions. Table 5-2 lists low pass filters' coefficients that are used for decimating and interpolating frames, where MPEG filter with 13 taps and AVC filter with 6 taps are used for decimation and interpolation processes respectively.

Table 5-2 Low pass filters coefficients used for decimation and interpolation

MPEG Filter	{2, 0, -4, -3, 5, 19, 26, 19, 5, -3, -4, 0, 2} / 64
AVC Filter	{1, -5, 20, 20, -5, 1} / 32

5.1.3 Experimental Setup

The first two views in each multi-view video sequence are coded through JM 18.0. IPPP coding structure is used, where each frame belonging to base view is predicted from a recent temporal frame. Frames that belong to dependent view are predicted through recent temporal and neighbouring spatial frames as shown in Figure 5-5. This figure shows two inter-view prediction directions. First direction deploys inter-view prediction via FR reference frames while the second applies inter-view prediction by LR reference frames. One hundred frames in each view are coded by stereoscopic video coding, where the frames from both views are interleaved via time-first coding order (Chen et al., 2009b). All coding modes are enabled while symmetric quality is applied for both prediction directions. This avoids any compensation for the negative effect of sub-sampled and up-sampled reference frames. Therefore the results will not be biased toward different quality in mixed spatial-resolution stereoscopic video. Quantisation step sizes are adjusted to match MVC common test conditions as shown in Table 5-3 (Su et al., 2006).

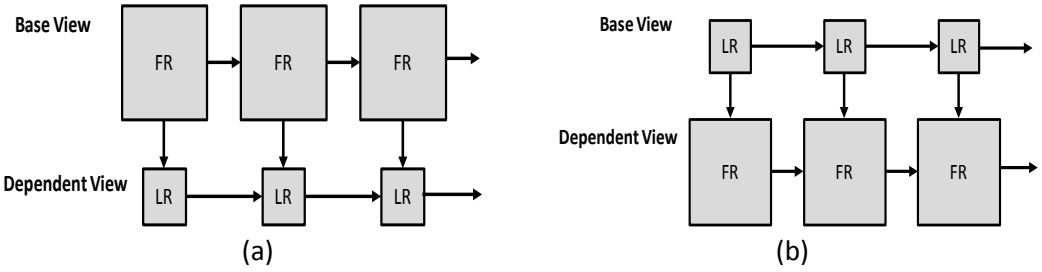


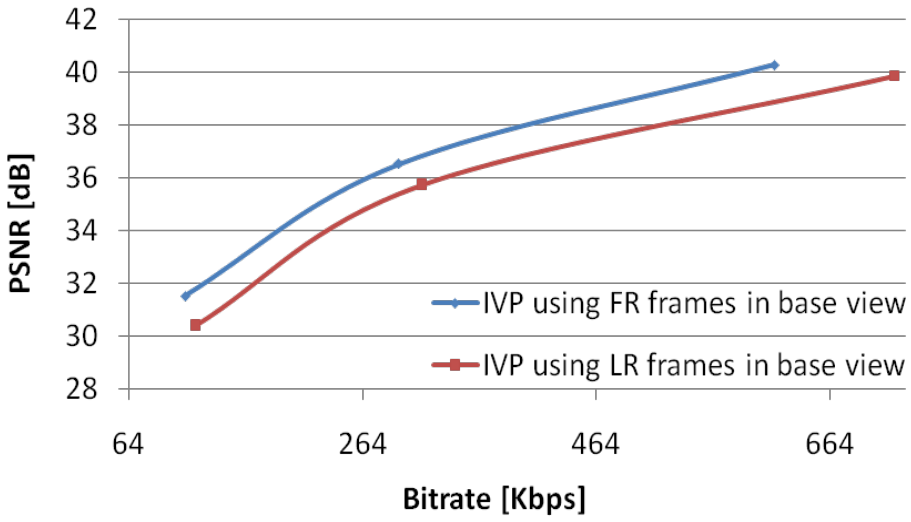
Figure 5-5 Mixed spatial-resolution stereoscopic video codec with two different inter-view prediction directions, where base view is a) FR and b) LR

Table 5-3 Quantisation parameter setting

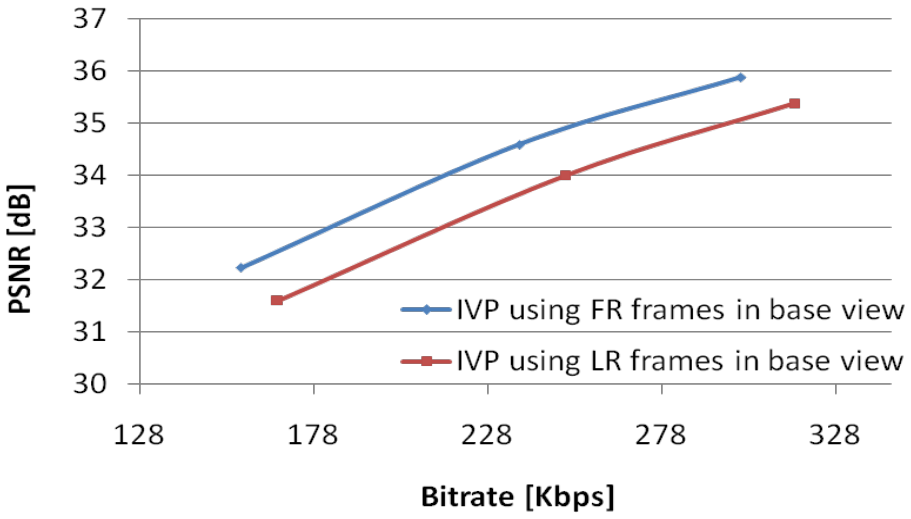
Sequence	High quality - QP_L	Medium quality -QP_M	low quality - QP_H
Break-dancers	22	26	31
Race1	24	26	28
Exit	26	29	31
Rena	23	28	33
Akko & Kayo	24	29	36
Ballroom	29	31	34

5.1.4 Results and Discussions

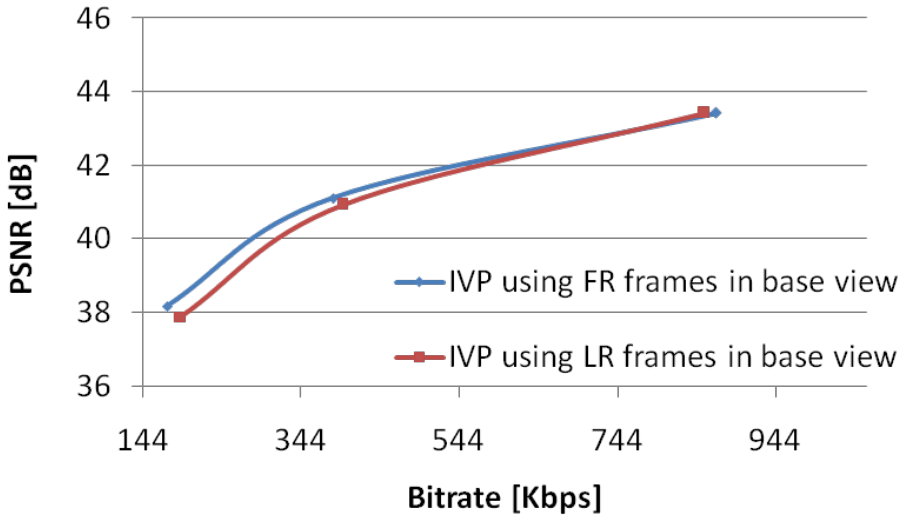
The coding performance for the two inter-view prediction directions is shown in Figure 5-6. It shows rate-distortion curves, where the horizontal and vertical axes are bitrate (Kbps) and *PSNR* (dB) respectively. Blue and red curves represent coding performance when FR or LR frames are used in base view.



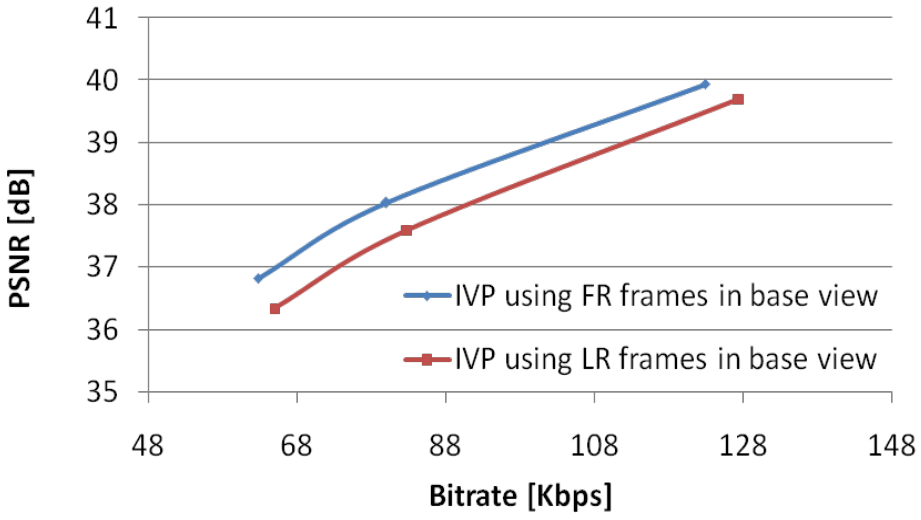
(a)


37
36
35
34
33
32
31
30
PSNR [dB]
IVP using FR frames in base view
IVP using LR frames in base view
128
178
228
278
328
Bitrate [Kbps]

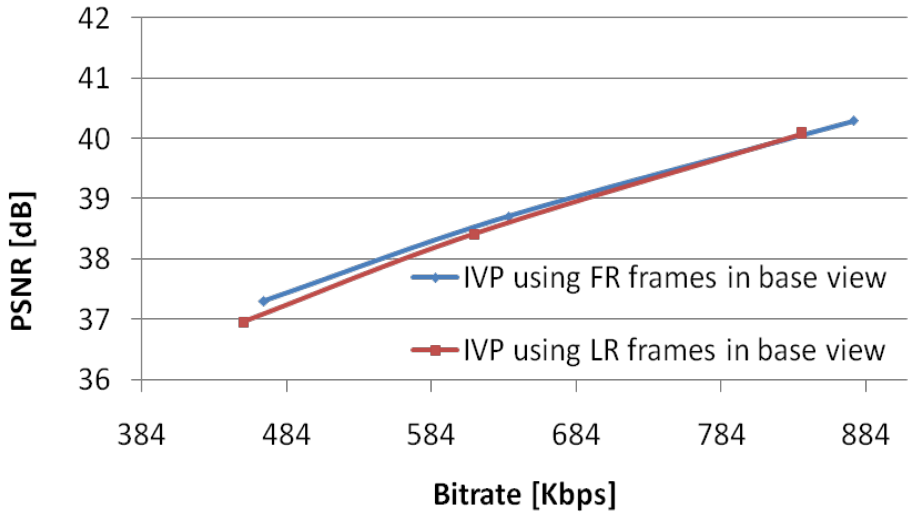
(b)


46
44
42
40
38
36
PSNR [dB]
IVP using FR frames in base view
IVP using LR frames in base view
144
344
544
744
944
Bitrate [Kbps]

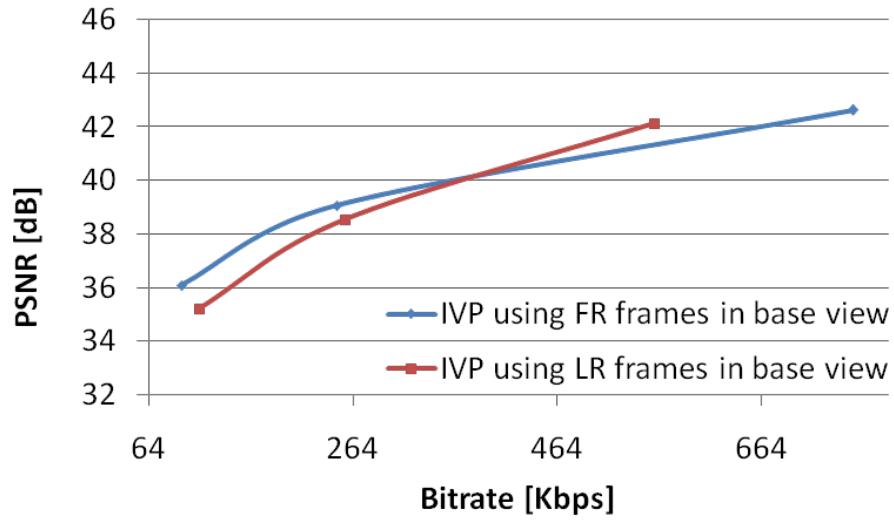
(c)


41
40
39
38
37
36
35
PSNR [dB]
IVP using FR frames in base view
IVP using LR frames in base view
48
68
88
108
128
148
Bitrate [Kbps]

(d)



(e)



(f)

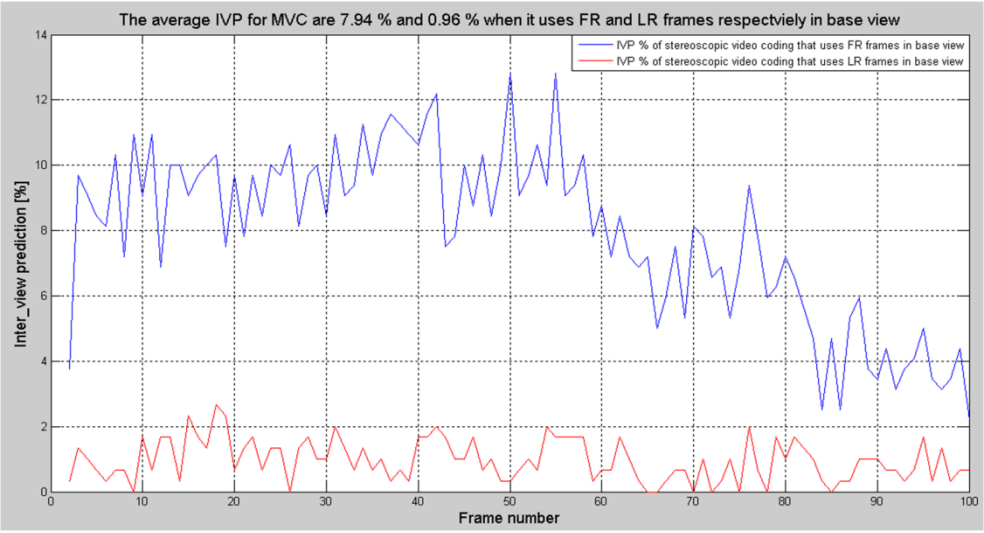
Figure 5-6 (a-f) Rate-distortion using mixed spatial-resolution stereoscopic video coding for Akko & Kayo, Ballroom, Break-dancers, Exit, Race1 and Rena respectively

From these Figures, mixed spatial-resolution stereoscopic video coding that uses FR frames as base view has higher coding performance than the corresponding codec that uses LR frames at low bitrates. The first IVP direction (blue curve) increases coding gain by on average 0.63 dB while it saves 6.2% of bitrate compared to IVP direction that uses LR frames in base view.

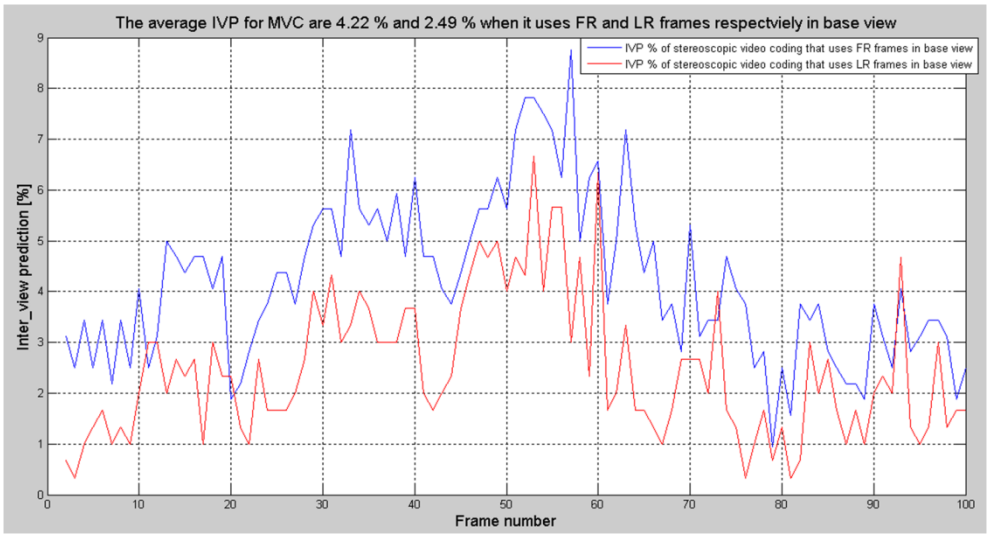
Mixed spatial-resolution stereoscopic video coding provides higher coding gain when FR frames are used in base view rather than LR frames. This is due to higher prediction accuracy that comes from neighbouring spatial frames. Statistical analysis for IVP has been applied at low bitrate. The amounts of blocks that are predicted via spatial frames that belong to base view in both IVP directions are compared. Figure

5-7 illustrates the amount of IVP in percentage across Y-axis for six stereoscopic videos while X-axis is the frame number that belongs to the dependent view. The average inter-view percentage in both prediction directions among every video is calculated. The ratio among both IVP directions is then obtained. The ratio is in the range of 1.2 to 8.8, where Break-dancers and Akko & Kayo videos have the lowest and highest ratios respectively.

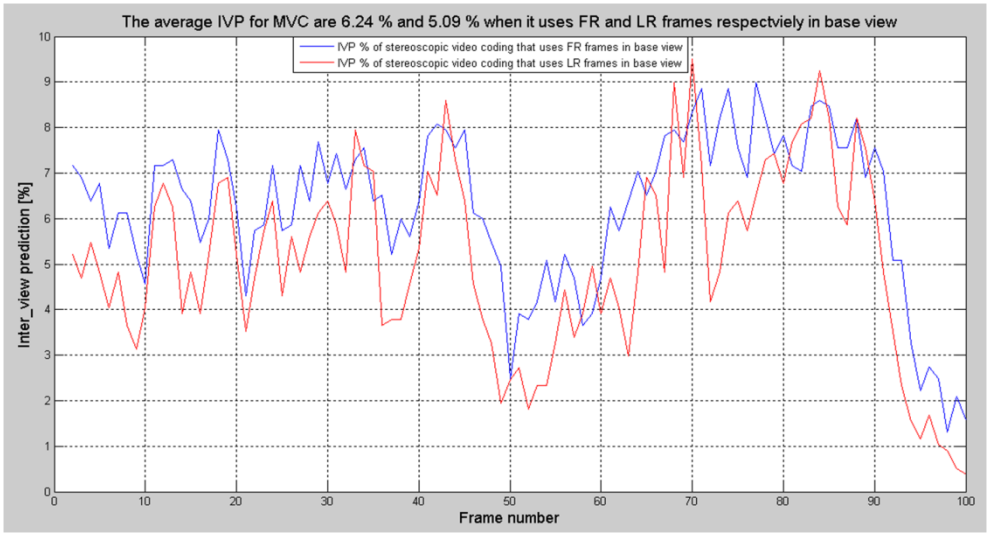
The coding performance for mixed spatial-resolution stereoscopic video codec depends on the type of frame resolution in base view. From the previous analysis results, deploying full rather than low spatial-resolution frames in the base view provides better inter-view prediction. When FR frames are used in base view, the reference frame has to be decimated to provide same spatial-resolution for the target frame that belongs to dependent view. Therefore decimated reference frame and target frame have similar information loss. However, the corresponding asymmetric coding that deploys LR frames in base view suffers from degradation in prediction efficiency. This is due to interpolation that is applied for the reference frame prior to prediction. The interpolated reference frame has blurriness which is strongly located around edges while the target frame maintains its high frequency (details), this would reduce the amount of inter-view prediction.



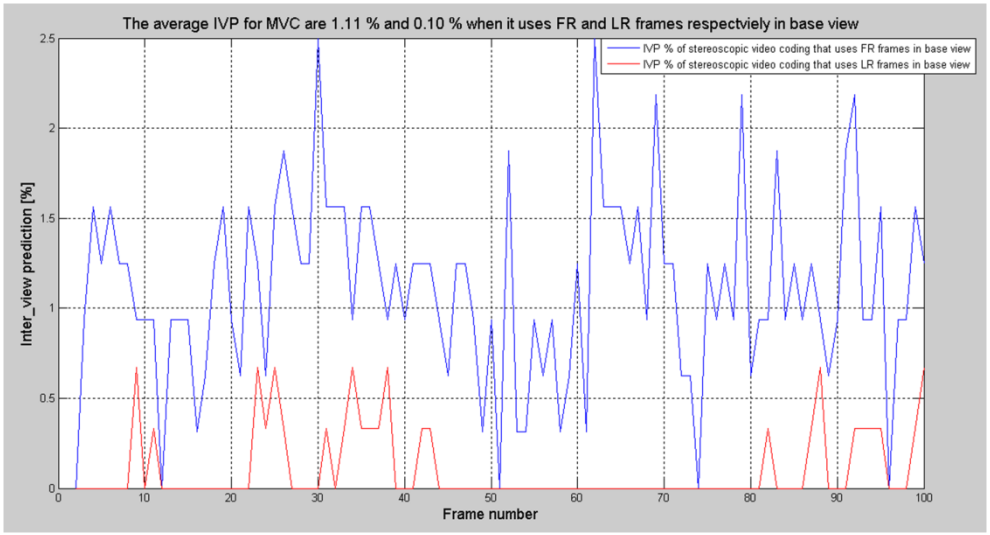
(a)


The average IVP for MVC are 4.22 % and 2.49 % when it uses FR and LR frames respectviely in base view
IVP % of stereoscopic video coding that uses FR frames in base view
IVP % of stereoscopic video coding that uses LR frames in base view
Inter_view prediction [%]
Frame number

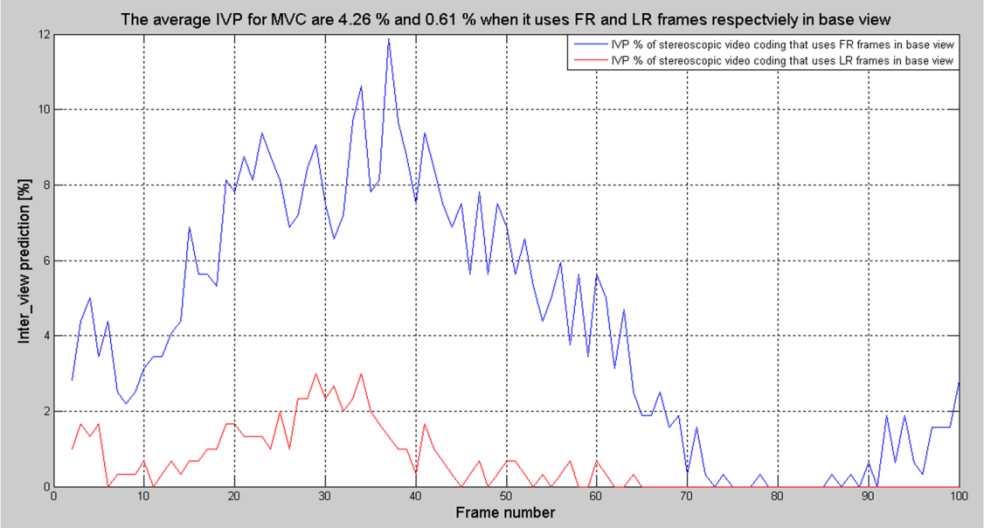
(b)


The average IVP for MVC are 6.24 % and 5.09 % when it uses FR and LR frames respectviely in base view
IVP % of stereoscopic video coding that uses FR frames in base view
IVP % of stereoscopic video coding that uses LR frames in base view
Inter_view prediction [%]
Frame number

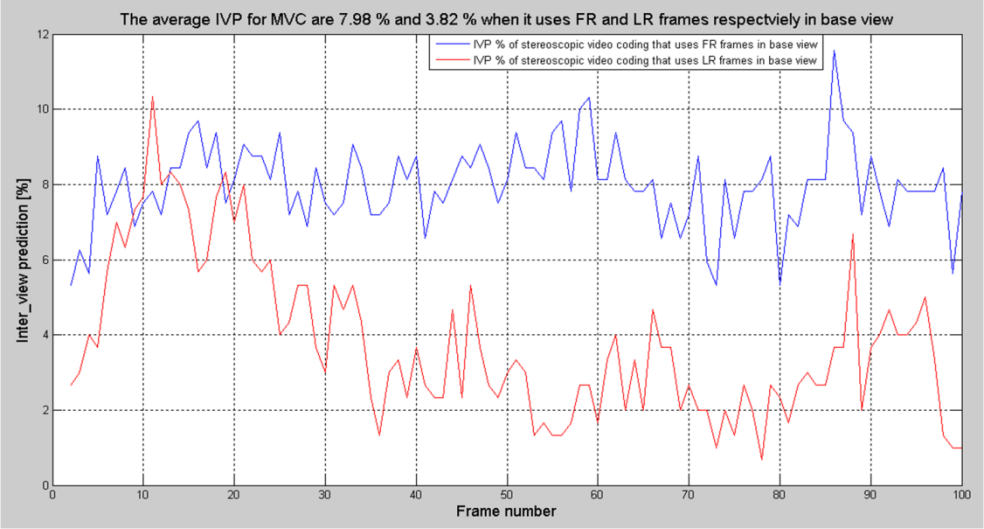
(c)


The average IVP for MVC are 1.11 % and 0.10 % when it uses FR and LR frames respectviely in base view
IVP % of stereoscopic video coding that uses FR frames in base view
IVP % of stereoscopic video coding that uses LR frames in base view
Inter_view prediction [%]
Frame number

(d)



(e)



(f)

Figure 5-7 (a-f) Amount of IVP for frames that belong to dependent view for Akko & Kayo, Ballroom, Break-dancers, Exit, Race1 and Rena respectively

Brust et al. investigated the effect of inter-view prediction direction on stereoscopic video coding (Brust et al., 2010). They reported similar coding efficiency for both inter-view prediction directions at low bitrates. Since their study considers asymmetric quality among mixed spatial-resolution stereoscopic videos, their results are biased to asymmetric quality setting. Delta quantisation for asymmetric coding that deploys LR frames (range of 7 to 9) is three times higher than the corresponding codec that uses FR frames in base view (range of 2 to 3). The following subsection will investigate the effect of asymmetric quality on inter-view prediction in the context of mixed spatial-resolution stereoscopic video coding.

5.1.5 Effect of asymmetric quality on the inter-view prediction

The relationship between inter-view prediction and quantisation parameter is explored in the context of mixed spatial-resolution stereoscopic video coding. Statistical analysis of block matching is deployed to reveal this relationship. Six videos are coded by H.264/AVC based stereoscopic video coding. These videos are Akko & Kayo, Ballroom, Break-dancers, Exit, Race1 and Rena. Low spatial-resolution frames are used in the base view while full spatial-resolution frames are used in the dependent view. Asymmetric quality has been applied such that delta quantisation (∆QP) values among neighbouring views are set from 0 to 10 with step size of 2. Low spatial-resolution frames are coded via lower QP than full spatial-resolution frames. Block matching statistics are analysed for blocks that are inter-view predicted using every quality setting.

Figure 5-8 shows the amount of inter-view prediction (%) when ∆QP changes for different stereoscopic videos, where ∆QP and inter-view prediction amount (%) are presented along X-axis and Y-axis respectively. From this figure, there is positive linear relationship among inter-view prediction and ∆QP parameter. Linear regression is used to analyse the data by using SPSS statistical software. Figure 5-9 shows curve fitting, where independent variable is ∆QP and dependent variable (response) is Inter-View Prediction (IVP). According to regression analysis results, there is a strong correlation[29] (0.665) among delta QP and IVP, where 44% of the variation in the inter-view prediction can be explained by asymmetric quality. The relationship of IVP and ∆QP based on six multi-view videos would be described by the following equation;

$$IVP = 1.492 + 1.096\,\Delta QP \tag{5-1}$$

Although deploying large ∆QP among mixed spatial-resolution frames would increase amount of inter-view prediction, it increases the amount of blockiness for full spatial-resolution frames that are coded by large QP (low-quality). Blocking artefacts degrades 3D perception since it is visible when low quality view is less than the threshold; approximately 32 dB (De Silva et al., 2012; Gurler & Tekalp, 2013). Therefore, asymmetric quality is a great challenge in the context of visual perception especially when the codec operates at low bitrates.

[29] Strong correlation for 0.665 is based on the interpretation described by Evans (Evans, 1996)

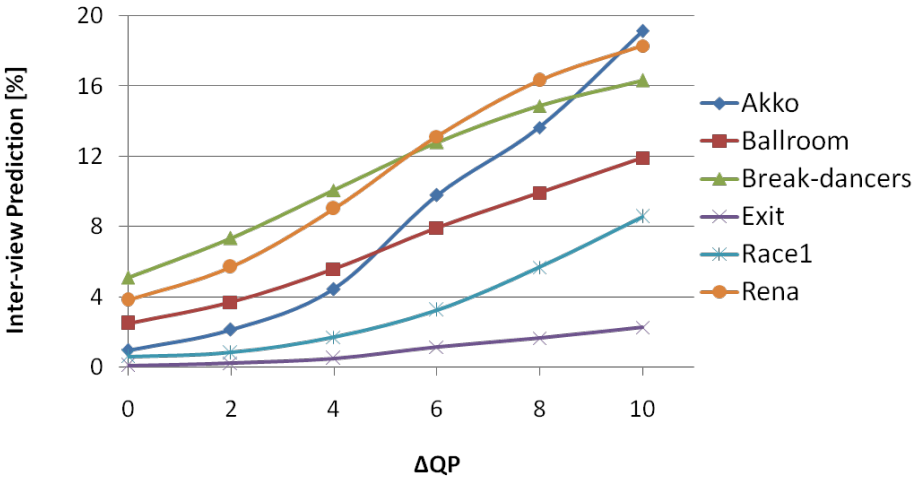


Figure 5-8 Relationship among ΔQP and inter-view prediction in percent

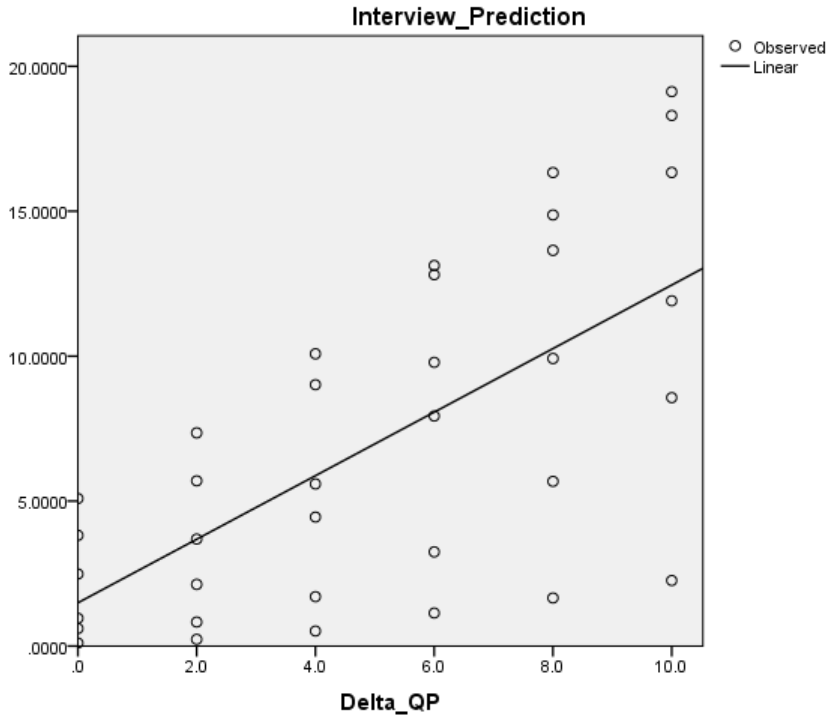


Figure 5-9 Curve fitting among delta quantisation parameter and IVP

5.1.6 Conclusions

This section investigated the effect of different inter-view prediction direction on the coding performance of mixed spatial-resolution stereoscopic video coding. At low bitrate, mixed spatial-resolution stereoscopic video coding provides superior coding performance using FR frames rather than LR frames in base view. When FR frames are used in base view, the reference frame and target frame have similar information loss due to decimation. However, the corresponding asymmetric coding that deploys

LR frames in base view suffers from degradation in prediction efficiency. This is due to blurriness that is consequence of interpolating low spatial-resolution reference frame, while the target frame maintains its high frequency (details); this would affect negatively the amount of inter-view prediction. The results obtained by *Brust et al.* are affected by asymmetric quality settings.

Mixed spatial-resolution multi-view video coding deploys decimating and interpolating reference frames in order to provide the same spatial-resolution for disparity estimation. The method used in decimation or interpolation is vital since it will affect integer and sub-pixel samples of reference frames. Therefore, the next section will examine different methods for decimating and interpolating reference frames. The objective is to find a suitable method for each process in terms of coding gain and computational complexity.

5.2 Different decimation and interpolation methods

5.2.1 Introduction

Decimation and interpolation are inevitable processes, where spatial-resolution for reference frame and target frame has to be the same prior to block matching. Decimation and interpolation are deployed at both encoder and decoder sides. Encoder needs to provide reference frames with the same spatial-resolution to current (target) frames during disparity estimation, while decoder performs these processes to decode current frames during disparity compensation. Since reference frame has sixteen samples, finding a suitable method for each process is important to reduce computational complexity overhead. The next two subsections will explore suitable methods for decimating and interpolating reference frames in terms of coding gain and computational complexity.

5.2.2 Different methods for decimating reference frames

This subsection explores different methods for decimating reference frames. Figure 5-10 provides an illustration for inter-view prediction among mixed spatial-resolution frames, where FR reference frame (F_1') is low pass filtered and down-sampled prior to disparity estimation of the target frame (F_2).

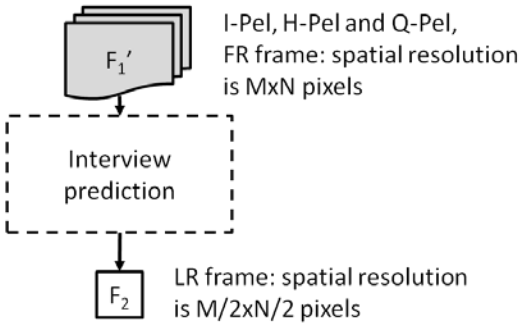


Figure 5-10 Inter-view prediction using FR reference frame

There are two methods for decimating reference frames. The first method (conventional decimation method) is widely used among mixed spatial-resolution stereoscopic video coding. This method filters and down-samples each sample separately (grey rectangles in Figure 5-10 represent F_1' samples). Figure 5-11-a shows this method, where LPF and head-down arrow stand for low pass filtering and down-sampling respectively. *Aflaki et al.* proposed high performance decimation method that is depicted in Figures 5-11-b (Aflaki et al., 2013b). High performance method gets first integer-pixel sample (I-pel) from the corresponding sample (FR). Half-pixel (H-pel) and Quarter-pixel (Q-pel) samples at LR are obtained from integer and sub-pixel samples at low spatial-resolution.

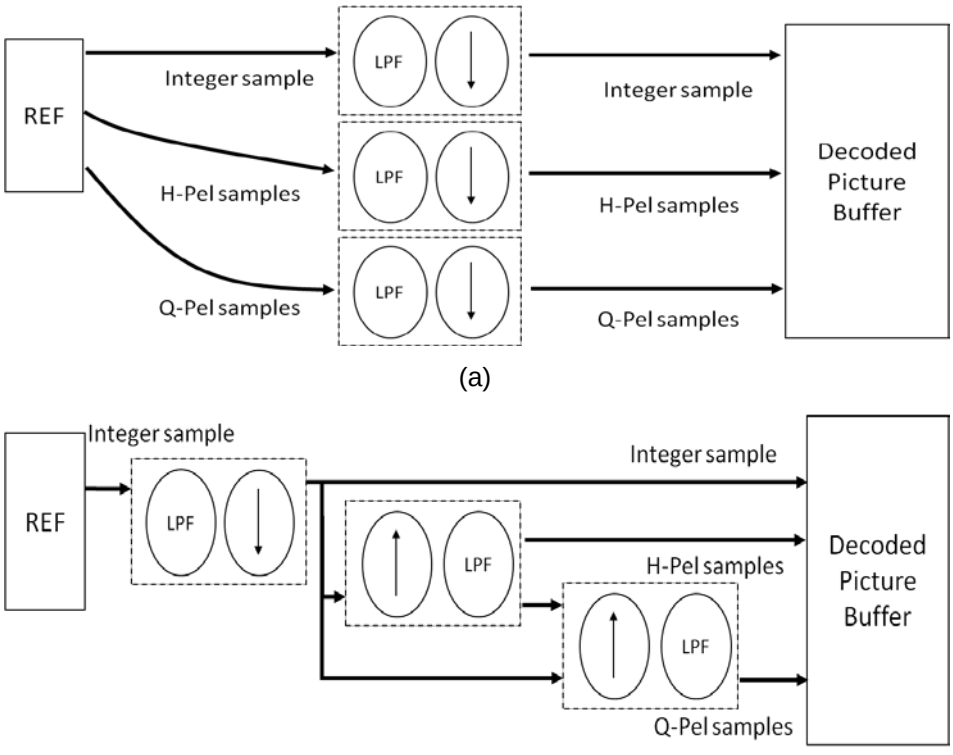
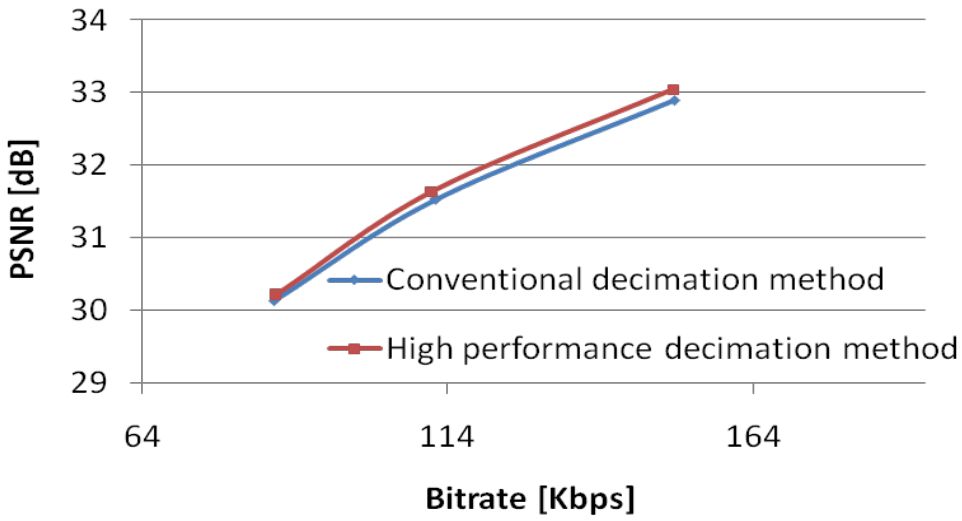


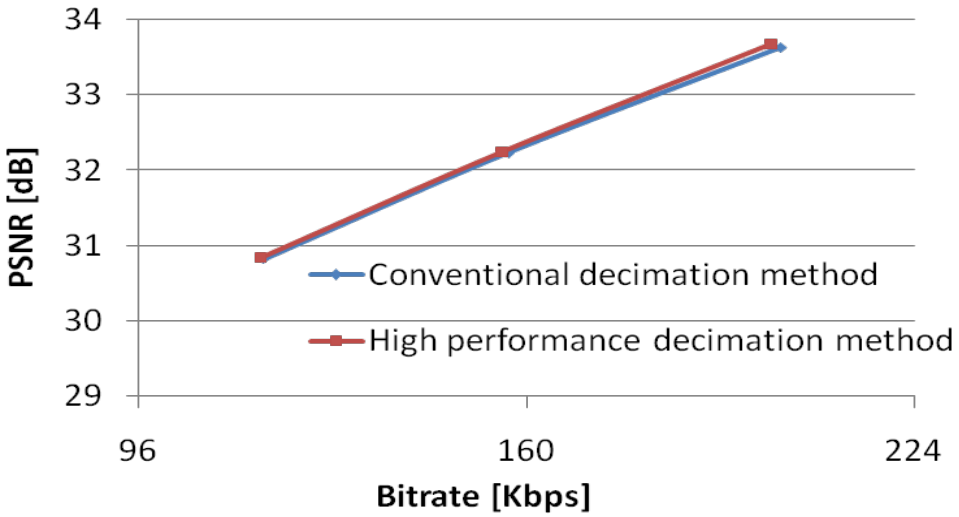
Figure 5-11 Decimation methods a) conventional and b) high performance method

Six mixed spatial-resolution stereoscopic videos are coded at low bitrates via H.264/AVC based stereoscopic video coding. These datasets are recommended by common test conditions for MVC (Su et al., 2006). Each video contains two-hundred frames, where base view uses FR frames. Full spatial-resolution reference frames are decimated before deploying disparity estimation. Two decimation methods are applied, where their consumed time during decimation and, coding performance are reported. Figure 5-12 shows rate-distortion curves for stereoscopic video coding. Conventional and high performance methods are presented using blue and red curves respectively. Bitrate (Kbps) and *PSNR* for luminance component (dB) are presented across X-axis and Y-axis respectively. From these figures, high performance method has slightly better coding performance than conventional method by saving bitrate on average by 0.88 Kbps.

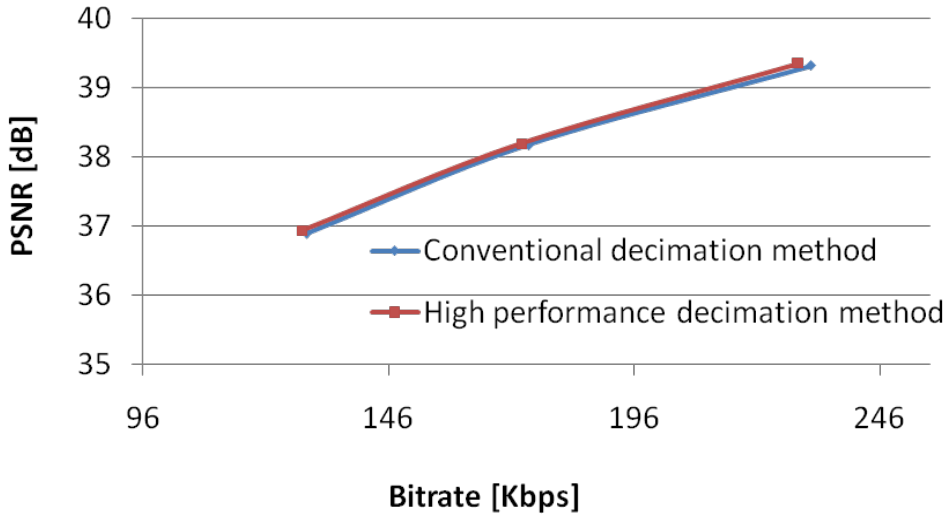
Figure 5-13 shows the total time consumed using these decimation methods. The measurements reflect the amount of computational complexity for each decimation method. All the experiments were carried out on a computer with Intel i7 CPU and memory of 16 GB. The total time consumed during coding Break-dancers is different than other datasets, as its frame's spatial-resolution is bigger than other datasets by a factor of 2.56. The average time required for conventional and high performance decimation methods are 112 seconds and 90 seconds respectively.



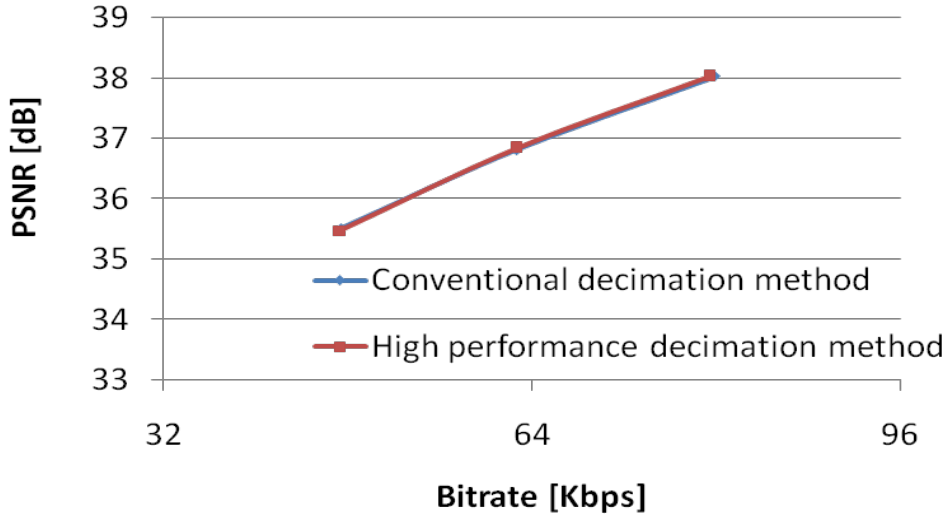
(a)


34
33
32
31
30
29
PSNR [dB]
Conventional decimation method
High performance decimation method
96
160
224
Bitrate [Kbps]

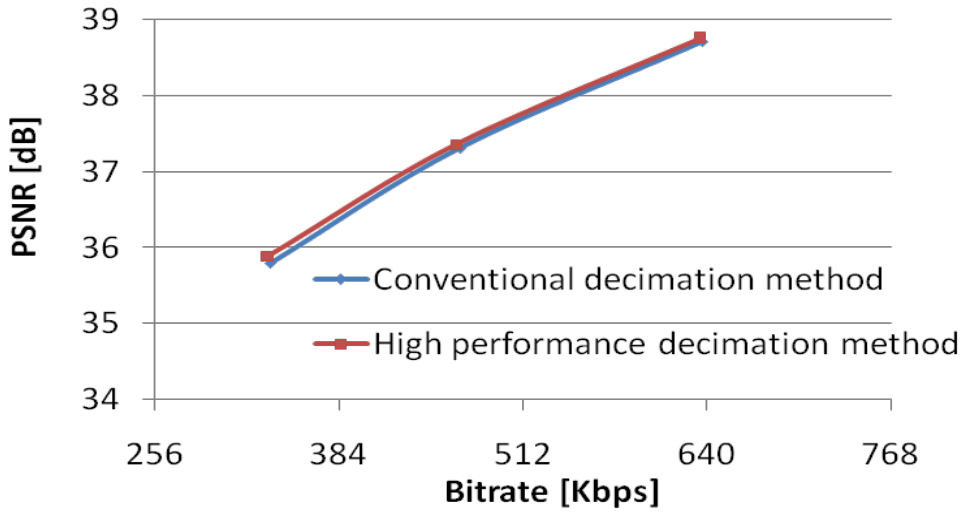
(b)


40
39
38
37
36
35
PSNR [dB]
Conventional decimation method
High performance decimation method
96
146
196
246
Bitrate [Kbps]

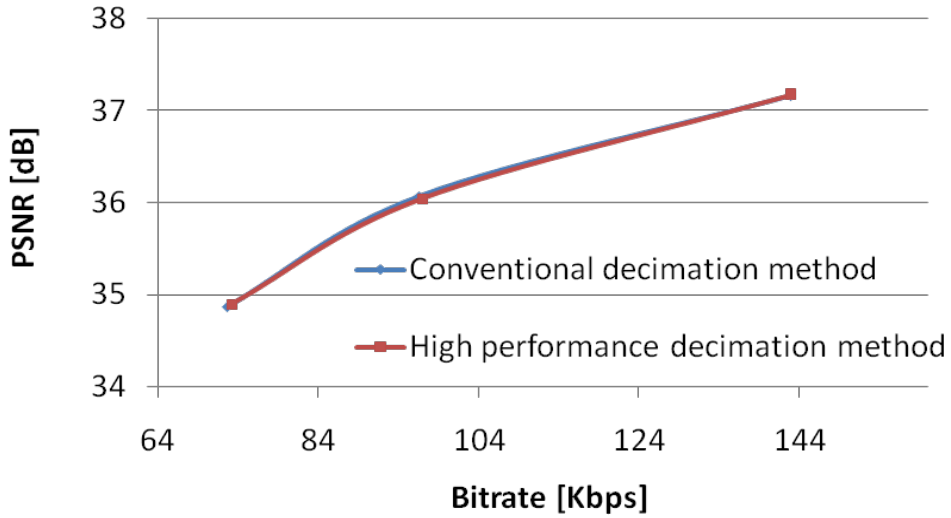
(c)


39
38
37
36
35
34
33
PSNR [dB]
Conventional decimation method
High performance decimation method
32
64
96
Bitrate [Kbps]

(d)



(e)



(f)

Figure 5-12 (a-f) Rate-distortion using different decimation methods for Akko & Kayo, Ballroom, Break-dancers, Exit, Race1 and Rena respectively

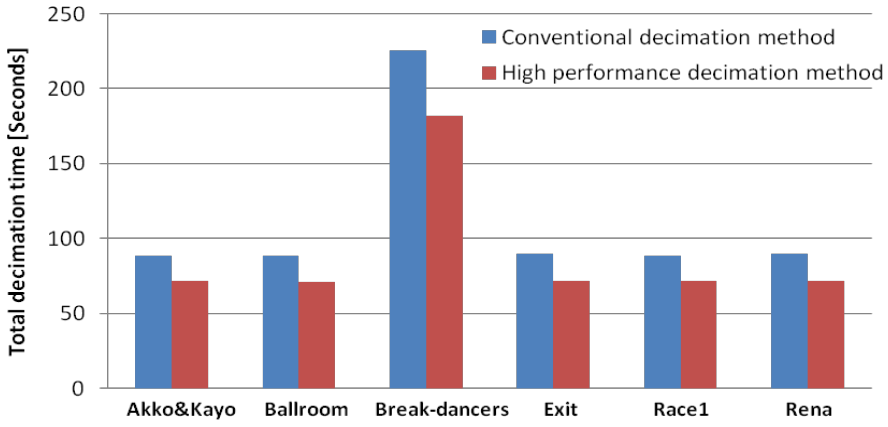
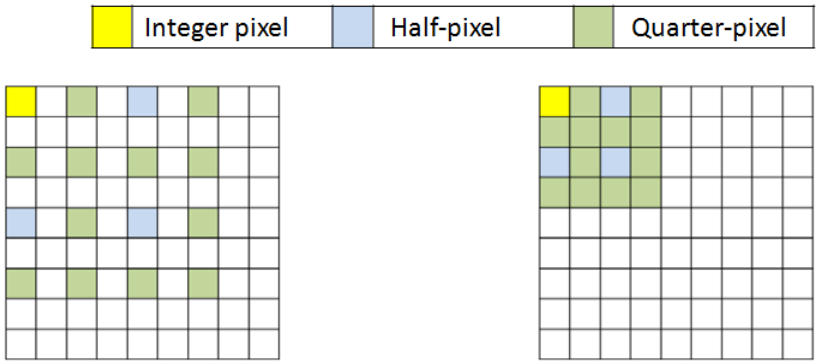


Figure 5-13 Total time consumed during decimating reference frames

High performance method is suitable for decimation in terms of coding gain and computational complexity. This method decreases the time needed for decimation by 24% compared to conventional decimation method without degrading quality of reference frames. The main part of the decimation complexity is the number of filter coefficients used, where large number of filter coefficients increases the amount of time needed for filtering the reference frame. The conventional method uses the same low pass filter (eleven non-zero coefficients) for sixteen samples. The high performance method applies this filter to obtain integer samples. It uses AVC interpolation filter (6 non-zero coefficients) three times to get half-pixel samples, while bilinear filter is used for quarter-pixel samples.

The high performance method distributes samples in a different way than the conventional method as depicted in Figure 5-14. Yellow, blue and green represent integer-pixel, half-pixel and quarter-pixel respectively. The filtered sub-pixels have uniform distribution when high performance method is used as shown in Figure 5-14-a. The filtered sub-pixels are localised in the first quadrant as shown in Figure 5-14-b when conventional method is applied. This results in high degree of similarity among samples when conventional method is used while samples obtained by high performance have less similarity. In order to realise the similarity among integer pixel and its sub-pixels, average Sum of Square Error (SSE_{avg}) is computed by using both methods for the first frame that belong to Akko & Kayo. Integer-pixel and its sub-pixels are used to calculate SSE_{avg} for luminance component. Table 5-4 shows SSE_{avg}, when high performance and conventional decimation methods are applied to first FR reference frame. From this table, samples have more similarity (less SSE_{avg}), when the conventional method is applied rather than the high performance method.



(a) (b)

Figure 5-14 (a-b) Integer and sub-pixels that represent reference frame samples using high performance and conventional decimation methods respectively

Table 5-4 SSE_{avg} for luminance component using high performance and conventional decimation methods

SSE_{avg} among all samples with respect to integer sample (I-Pel)	FR reference frame is decimated by high performance method				FR reference frame is decimated by conventional method			
	I-Pel	**Q-Pel**	**H-Pel**	**Q-Pel**	**I-Pel**	**Q-Pel**	**H-Pel**	**Q-Pel**
I-Pel	0	14.4	57.3	115.5	0	3.1	11.9	25.5
Q-Pel	26.92	41.1	80.9	136.4	5.5	8.2	16.6	29.8
H-Pel	107.4	118.4	158.5	212	21.1	23.5	31.8	44.7
Q-Pel	217.4	225.1	263.3	317.3	45.7	47.6	55.6	68.2

5.2.3 Different methods for interpolating reference frames

Low spatial-resolution frames in base view need to be interpolated prior to disparity estimation as shown in Figure 5-15. Interpolating reference frames does not bring new information or cause information loss. The low spatial-resolution frame is up-sampled by a factor of two in the horizontal and vertical directions. The padded pixels are then generated via low pass filter.

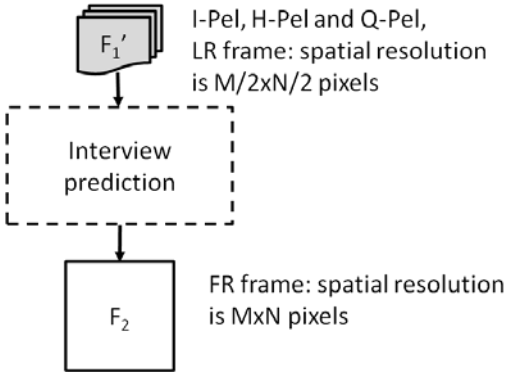


Figure 5-15 Inter-view prediction using LR reference frame

The conventional interpolation method handles each sample separately, where each is up-sampled and filtered via AVC interpolation filter (6-taps). The method is illustrated in Figure 5-16-a. The second method is high performance interpolation method that is opposite to the corresponding high performance decimation method. It interpolates integer-pixel sample using corresponding low spatial-resolution reference frame. This integer sample is used to estimate half-pixel samples. The remaining samples are obtained via integer-pixel and half-pixel samples that belong to full spatial-resolution frame as shown in Figure 5-16-b.

Six mixed spatial-resolution stereoscopic videos are coded, where each has two-hundred frames. Base view has LR frames while dependent view has FR frames. Two interpolation methods are compared in terms of coding gain and computational

complexity. Figure 5-17 shows rate-distortion curves for stereoscopic video coding, where blue and red curves are coding performances using conventional and high performance methods respectively. Bitrate (Kbps) and *PSNR* for luminance component (dB) are presented across X-axis and Y-axis respectively. From these figures, conventional and high performance interpolation methods give the same coding performance. Figure 5-18 shows total time consumed during interpolating reference frames, where UP refers to up-sampling. All the experiments were carried out on a computer with Intel i7 CPU and memory of 16 GB. From this figure, interpolating samples using the high performance method reduces the amount of time needed for interpolation up to 56% with respect to time needed by the conventional method. AVC interpolation filter (6-coefficients) is used by sixteen and four times when the conventional and high performance interpolation methods are used respectively. Hence, the conventional method consume more time for filtering than the high performance method.

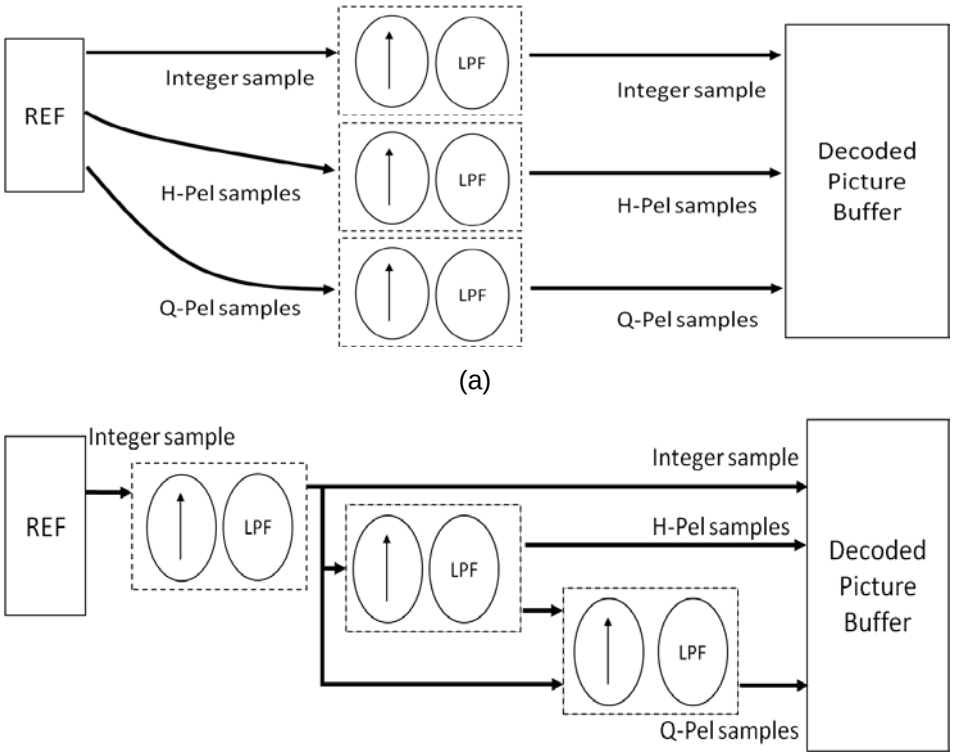
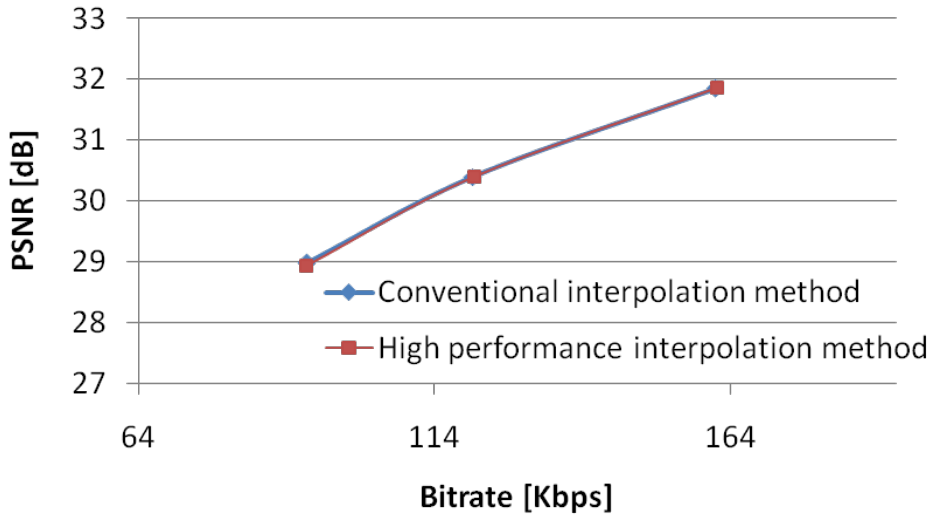
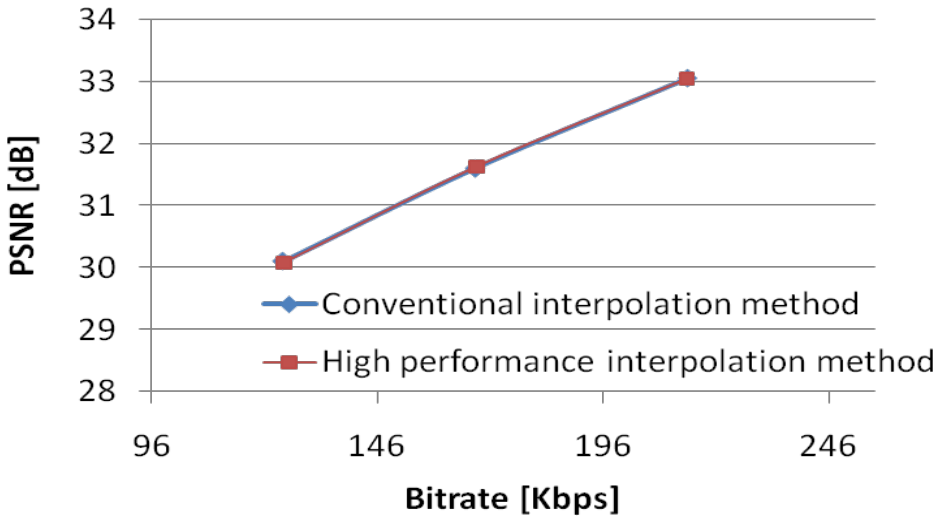


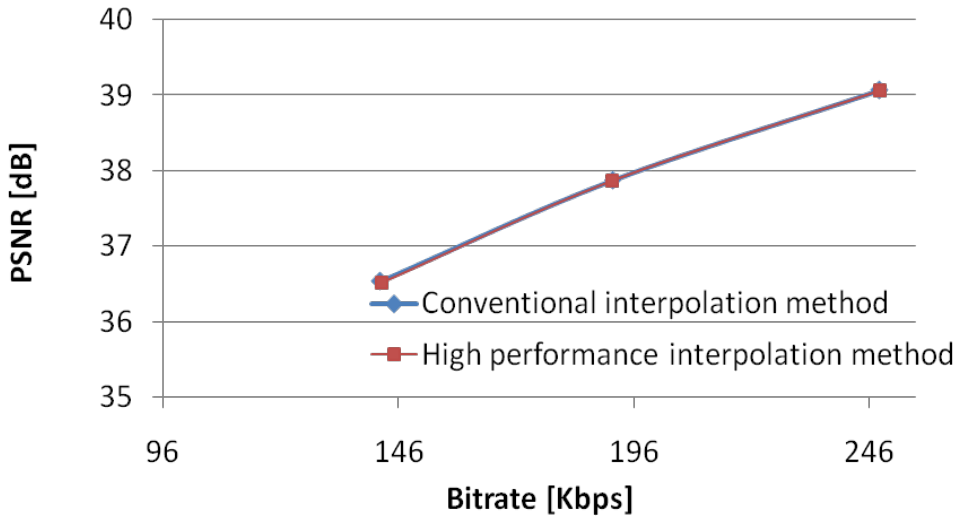
Figure 5-16 (a-b) Reference frame interpolation using conventional and high performance methods respectively


33
32
31
30
29
28
27
PSNR [dB]
Conventional interpolation method
High performance interpolation method
64
114
164
Bitrate [Kbps]

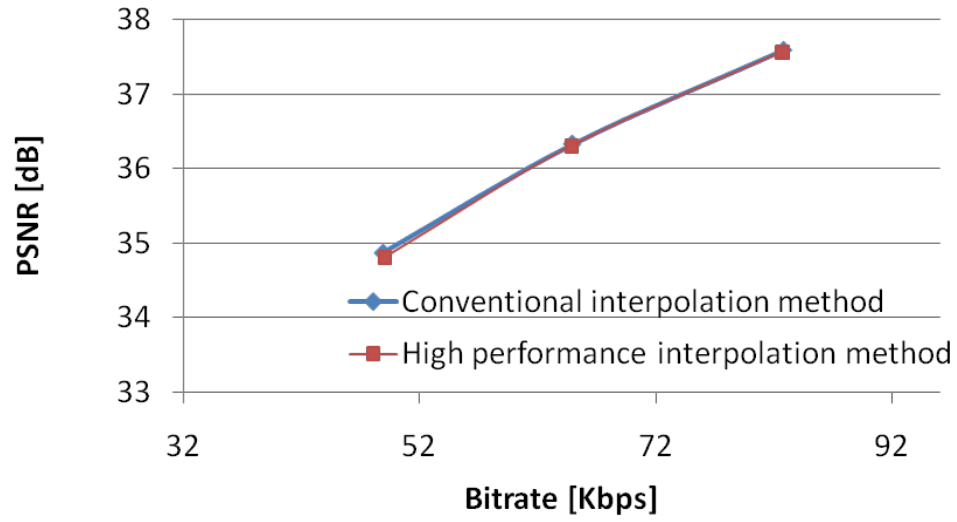
(a)


34
33
32
31
30
29
28
PSNR [dB]
Conventional interpolation method
High performance interpolation method
96
146
196
246
Bitrate [Kbps]

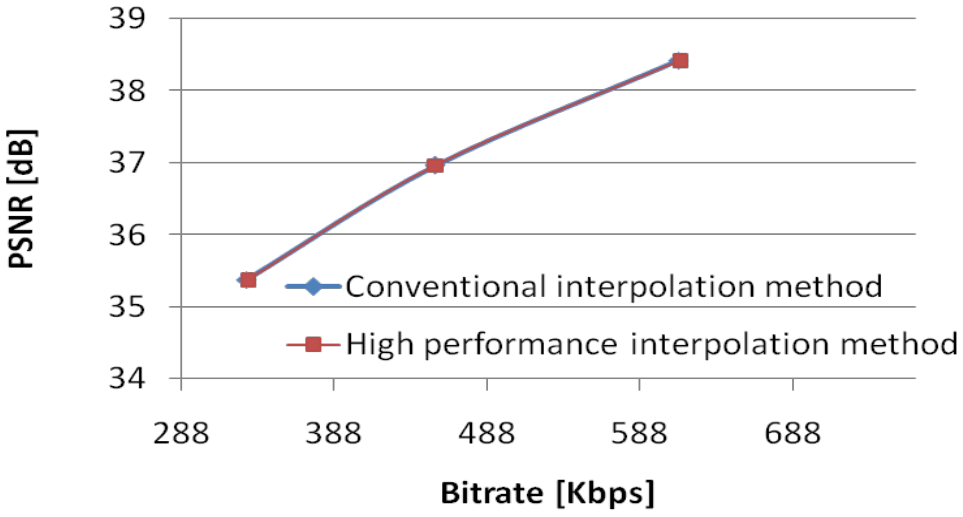
(b)


40
39
38
37
36
35
PSNR [dB]
Conventional interpolation method
High performance interpolation method
96
146
196
246
Bitrate [Kbps]

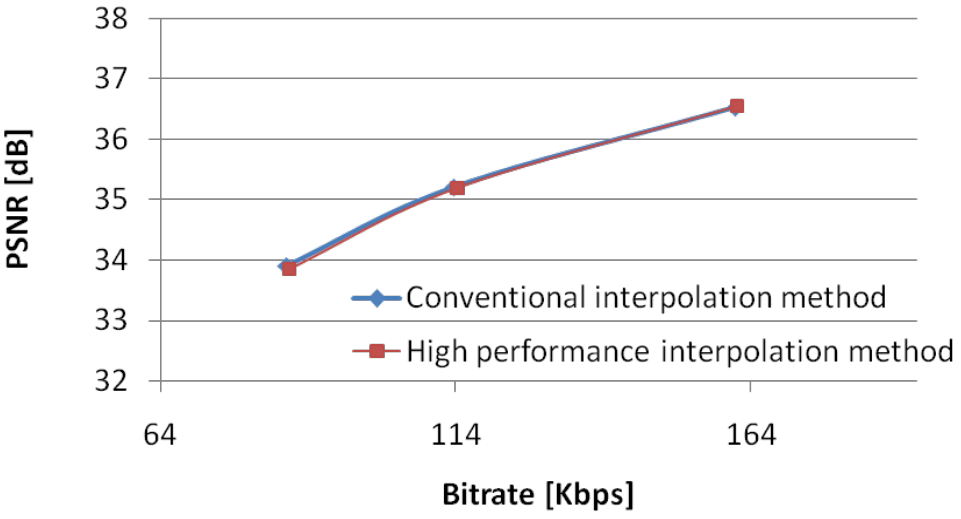
(c)



(d)



(e)



(f)

Figure 5-17 (a-f) Rate-distortion using different interpolation methods for Akko & Kayo, Ballroom, Break-dancers, Exit, Race1 and Rena respectively

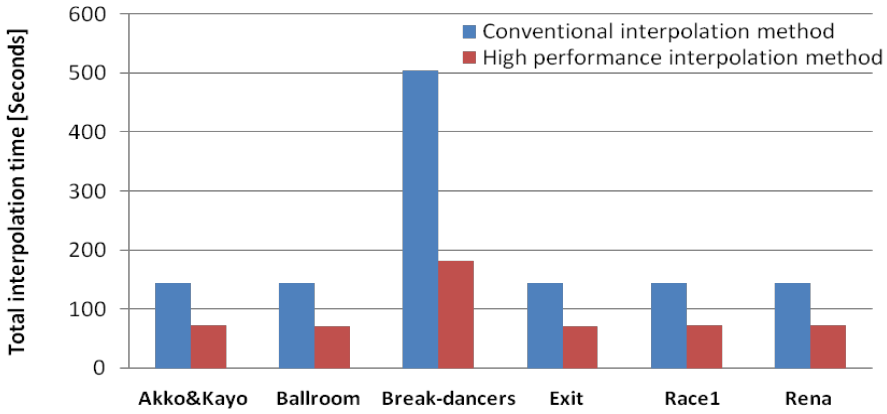


Figure 5-18 Total time consumed during interpolating reference frames

5.2.4 Conclusions

This section investigated different methods for decimating and interpolating reference frames. Conventional methods for decimation and interpolation would cost significant amount of time, where each sample is filtered separately. High performance methods reduce the amount of time needed for decimation and interpolation through filtering fewer numbers of samples. According to coding performance and time needed for filtering, the high performance methods are recommended for decimation and interpolation. Disadvantage from relying on high performance method is removing the one-to-one relationship among samples at low and full spatial-resolution. Conventional method maintains this relationship, where each sample depends on the corresponding sample presented at different spatial-resolution.

The following section investigates the prediction architectures for mixed spatial-resolution multi-view video coding. Statistical analysis of block matching among candidate reference frames will be used to derive prediction architecture.

5.3 Mixed spatial-resolution multi-view video coding using statistics of block matching

5.3.1 Introduction

This section investigates prediction architectures for mixed spatial-resolution MVC. Through block matching analysis among neighbouring reference frames, RFS and RFO will be defined. Two block matching statistical analyses are applied separately

for both, full and low spatial-resolution frames. Since the contribution of reference frames change with time, prediction via spatial and temporal reference frames should be avoided when the expected amount of block matching for each frame is insignificant. Therefore, a study is conducted to investigate the feasibility of dynamically skipping these reference frames. The proposed prediction architecture will then be presented and evaluated among other prediction architectures in terms of coding performance, computational complexity at the encoder side and memory consumption at the decoder side. Adaptive reference frame ordering algorithm (presented in section 4.4) will be integrated with the proposed architecture. The outcomes will be concluded at the end of this section.

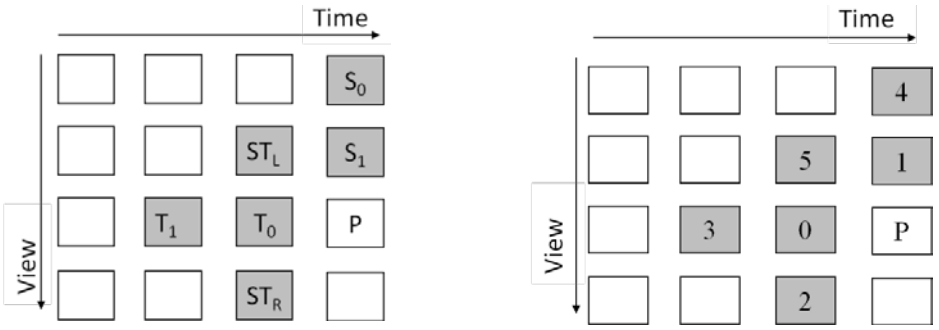
5.3.2 Statistics of block matching among reference frames

The outcomes from the previous block matching statistics (discussed in section 4.3) are used to preliminary define RFS that is deployed in this subsection. Accordingly, six reference frames; T_0, T_1, S_0, S_1, ST_L and ST_R are used in the analysis as shown in Figure 5-19. Two separate block matching statistics are analysed for Break-dancers. Four successive views are coded via H.264/AVC based multi-view video coding, where different combinations are coded. Five sequences[30] are examined; their output results are averaged to reveal the block matching contribution for each reference frame. The reference frame ordering for each analysis is depicted in Figure 5-20. Based on the results derived from the first section (5.1); FR reference frames indices are placed first in List 0 when coding FR frames as shown in Figure 5-20-a. Predicting LR frames via full spatial-resolution reference frames do not negatively affect inter-view prediction. Therefore, RFO follows the corresponding order for symmetric multi-view video coding (Figure 5-20-b) when conducting block matching analysis for LR frames.

Table 5-5 shows the average results for Block Matching (BM) statistics in percentage among reference frames when coding five sequences. From this table, two most significant reference frames for predicting full spatial-resolution frame are T_0 and S_0. These frames contribute by on average 91.1%. For low spatial-resolution frame, both T_0 and S_1 have significant role of block matching that is on average 92.2%. Coding LR frames provide easier scenario to define RFS than FR frames. When coding LR frames, two closest reference frames (temporal and spatial) provide the majority of block matching with respect to the remaining reference frames. Predicting FR frames are affected negatively by the neighbouring reference frames that have lower spatial-resolution. This increases the amount of blocks that are

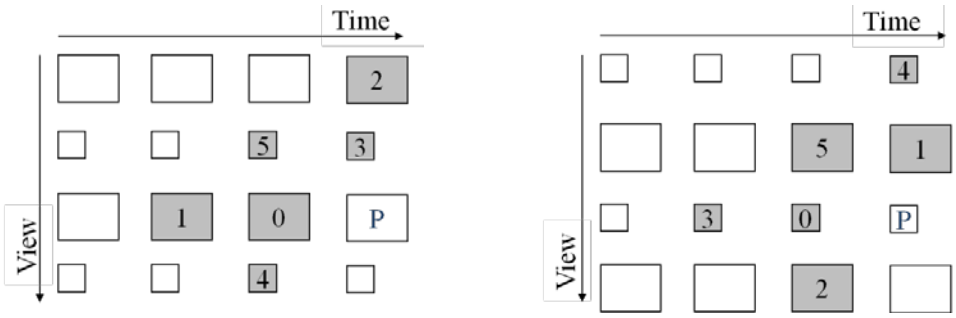
[30] Starts from sequence that includes view 0 to 3 until last sequence that has view 4 to 7

temporal predicted (83.3%). Spatiotemporal reference frames provide less than 1% when predicting FR frames while they contribute by more than 5% for LR frames. Figure 5-21 presents RFS and RFO when number of reference frames is two, where the number inside each block is the reference frame order. Inter-view prediction is mostly affected by blurred reference frames when coding FR frames. From Table 5-5, neighbouring FR reference frame provides higher block matching than the closest LR frame.



(a) (b)

Figure 5-19 Symmetric multi-view video coding a) RFS and b) RFO



(a) (b)

Figure 5-20 RFO for block matching statistics when coding a) FR and b) LR frames

Table 5-5 Statistical analyses average results when coding FR and LR frames

BM statistical analysis (%)	**T_0**	**T_1**	**S_0**	**S_1**	**ST_R**	**ST_L**
Full spatial-resolution frame	79.29	3.96	11.81	4	0.58	0.36
Low spatial-resolution frame	60.27	1.78	0.87	31.97	4.17	0.94

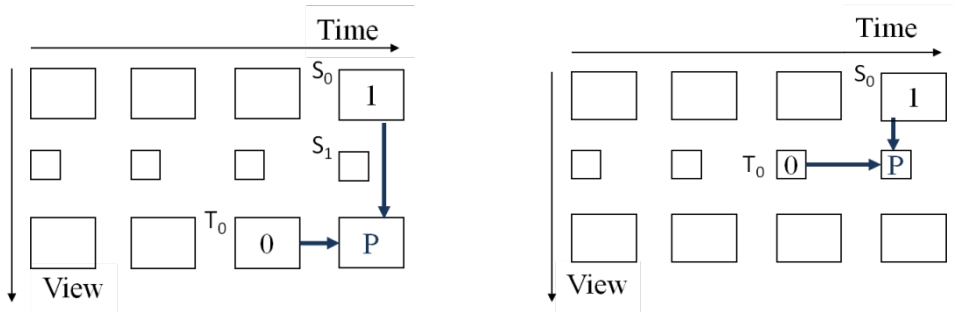


Figure 5-21 Reference frame selection and reference frame ordering for a) full and b) low spatial-resolution frames

The block matching statistics have been used to compare the amounts of inter-view prediction using both FR and LR frames (S_0 and S_1 frames as shown in Figure 5-22-a). H.264/AVC based MVC has been used to compress three-view video that deploys mixed spatial-resolution frames. The middle view contains LR frames, while surrounding views have FR frames. Six videos have been coded, where one-hundred frames from each view has been coded. Two prediction architectures have been deployed as shown in Figure 5-22-b and Figure 5-22-c. They use recent temporal and spatial reference frame in predicting FR frames that belong to the third view. The 1st PA uses FR frame (S_0), while the second PA uses neighbouring LR frame (spatial reference frame, S_1). The amount of IVP for FR frames that belong to the third view are analysed for both S_0 and S_1 reference frames, where the results are shown in Table 5-6. The outcomes from this table are consistent with the results presented in Table 5-5. Although LR frame (S_1) has more spatial redundancies than FR frame (S_0) with respect to current FR frame (P), it provides less significant role for block matching than S_0. This is due to high frequency components that exist in S_0 and P-frame, where both are not decimated prior to disparity estimation. High frequency components in LR frame are degraded that negatively affects its IVP accuracy. Therefore, FR reference frames should be used instead of LR frames in IVP for FR frames that belong to the dependent view.

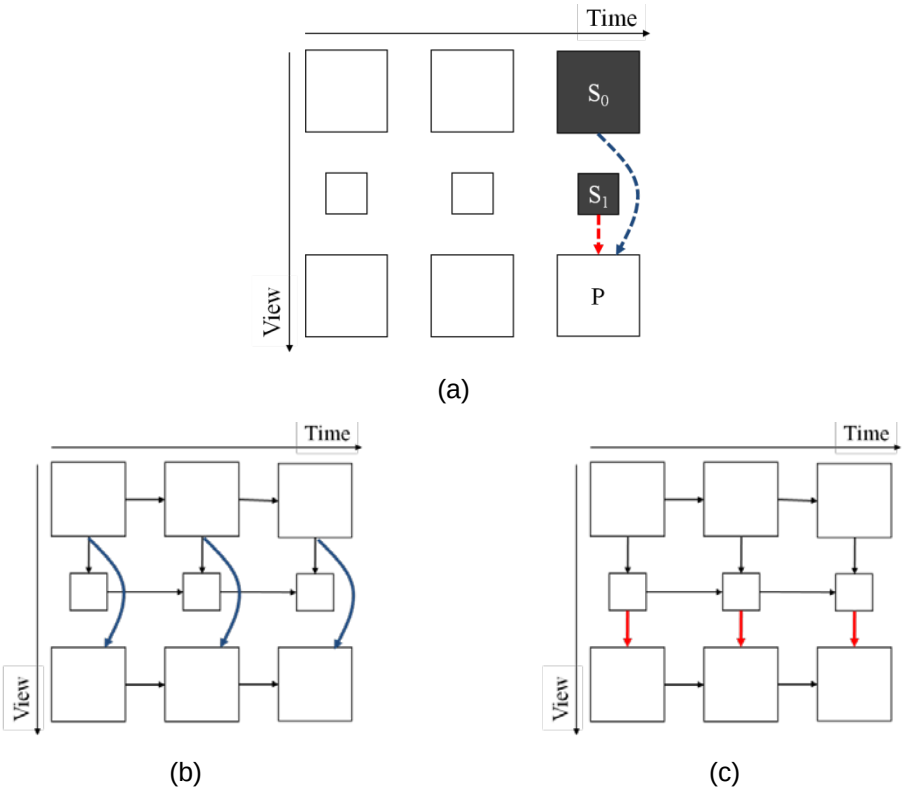


Figure 5-22 (a-c) shows different IVP sources, PA using FR and LR frames for IVP

Table 5-6 Average IVP amount (%) when FR frame is predicted using S_0 and S_1 frames

Average IVP (%) using	Akko & Kayo	Ballroom	Break-dancers	Exit	Race1	Rena	Average
FR frame (S_0)	9.18	8.75	16.58	1.08	7.76	9.14	8.75
LR frame (S_1)	4.54	6.77	14.4	0.72	3.81	8.73	6.5

Coding FR frames that belong to a dependent view is most challenging in prediction architecture, where the amount of IVP using S_0 and S_1 is 15.8 % compared to the corresponding amount (32.8%) for LR frames as shown in Table 5-5. Number of blocks in the full spatial-resolution frame is higher than the corresponding blocks in LR frame by factor of four. Therefore, second temporal frame is used for predicting FR frames that belong to dependent view. Multi-view videos have various characteristics in terms of scene complexity and objects motion. This affects the efficiency of temporal and inter-view predictions that consequently influences reference frame selection and reference frame ordering. The following subsection will investigate the feasibility of dynamically skipping these reference frames when the expected amount of block matching for each reference frame is insignificant during coding the full spatial-resolution frame.

5.3.3 Dynamic temporal and spatial reference frames selection

Multi-view video using three-views is explored, where low spatial-resolution frames are deployed in the middle view while full spatial-resolution frames are used in base and third views. Figure 5-23 shows three-view video, where RFS for frames that belong to dependent view use two recent temporal frames and FR spatial reference frame. Based on the results reported in subsection 5.3.2, spatial reference frame (S_0) is used to predict frames in neighbouring views (A and B frames) as shown in Figure 5-23. Therefore, the amount of inter-view predicted blocks in A-frame and B-frame might be correlated. This would entail the feasibility of exploiting the amount of inter-view predicted blocks in A-frame to dynamically select S_0 before coding B-frame. To validate this correlation, statistical analysis is applied to compute the amount of inter-view predicted blocks for A-frame and B-frame using S_0 as shown in Figure 5-24 that are referred to blue and red arrows respectively. The correlation among inter-view predicted blocks when coding low and full spatial-resolution frames is presented in Table 5-7. The average inter-view prediction correlation based on six videos is 0.44. It shows moderate[31] positive relationship among the amount of inter-view predicted blocks when coding LR and FR frames. Therefore, the amount of inter-view predicted blocks for LR frame (A-frame) is analysed. When this amount is less than threshold[32], then S_0 reference frame is skipped during coding FR frame (B-frame).

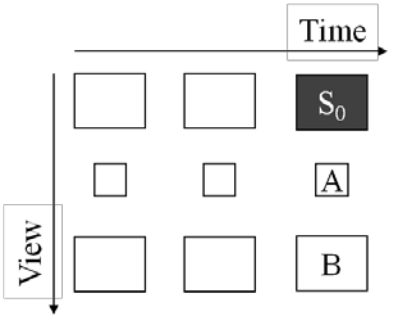
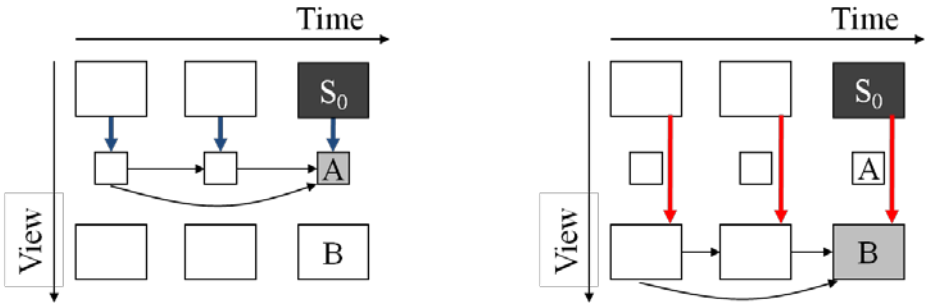


Figure 5-23 Inter-view prediction for LR and FR frames

[31] Moderate correlation for 0.44 is based on the interpretation described by Evans (Evans, 1996)
[32] It will be defined at the end of this subsection.



(a) (b)

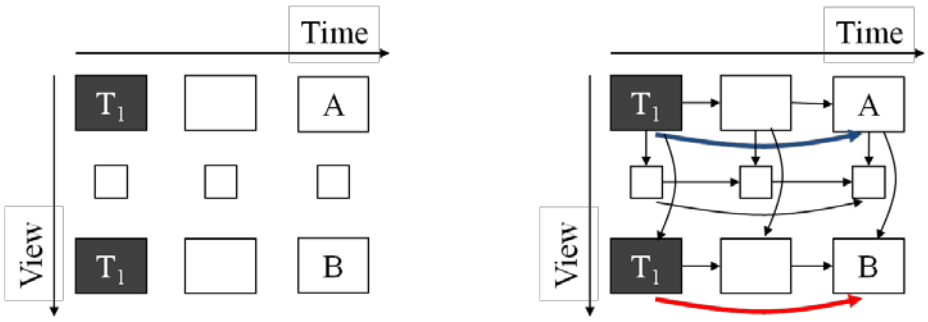
Figure 5-24 Inter-view prediction for a) LR frame and b) FR frame

Table 5-7 Average inter-view prediction correlation among LR and FR frames

Dataset	Akko & Kayo	Ballroom	Break-dancers	Exit	Race1	Rena	Average
Average correlation for IVP	0.16	0.57	0.74	0.22	0.66	0.3	0.44

The Second temporal frame (T_1) is used during coding FR frame that belong to dependent view when it is expected to significantly contribute to block matching. This is applicable when a correlation exists among temporal predicted blocks using 2nd temporal reference frame among frames that belong to base and third views as shown in Figure 5-25-a. Statistical analysis has been conducted in order to validate the correlation among temporal-predicted blocks in both frames; A-frame and B-frame. Figure 5-25-b highlights temporal reference frame; T_1, where its role of block matching is analysed when coding both FR frames (referred to T_1 via blue and red arrows). The correlation result is tabulated in Table 5-8. It shows moderate positive relationship[33] (0.42) among 2nd temporal reference frame during coding A-frame and B-frame. Based on the correlation results in Table 5-8, T_1 temporal reference frame is used during coding B-frame when the corresponding amount for coding A-frame is higher than the threshold. Since Exit MVV contains objects with slow motion characteristics, the amount of temporal predicted blocks during coding FR frames (A-frame and B-frame using T_1) is not significantly high (0.98% and 0.99% respectively). Although Exit MVV shows negative correlation among temporal prediction for FR frames that belong to the base and dependent views, the amount of correlation is very weak (0.08) as shown in Table 5-8. Therefore there is almost no correlation among the few blocks that are predicted using T_1 for this particular MVV.

[33] Moderate correlation for 0.42 is based on the interpretation described by Evans (Evans, 1996)



(a) (b)

Figure 5-25 a) Temporal prediction using 2nd temporal reference frame source and b) Prediction architecture among three-view video coding

Table 5-8 Average temporal prediction correlation among FR frames

Dataset	Akko & Kayo	Ballroom	Break-dancers	Exit	Race1	Rena	Average
Average correlation for temporal prediction	0.11	0.43	0.67	-0.08	0.82	0.57	0.42

Dynamic spatial and temporal reference frames selections are deployed during coding FR frames that belong to the third view. They are selected when the corresponding amount of inter-view and temporal predicted blocks are higher than the threshold. It would refer to insignificant amount of block matching. To set the threshold value, six videos have been coded via H.264/AVC based multi-view video coding, where different thresholds are used (0%, 2.5%, 4%, 6%, 12% and 20%). The thresholds have been chosen since four, six, twelve and twenty percent have been used in the literature to describe different amounts of block matching that reflect to very low, low, significant and high amounts of block matching respectively (Kaup & Fecker, 2006; Merkle et al., 2007a; Shen et al., 2007). Increasing threshold value reduces the amount of time needed to encode multi-view video through skipping more reference frames at the expense of increasing average bitrate with respect to the same codec that does not apply threshold. Figure 5-26 shows the effect of using different values of threshold on the bitrate. From this figure, setting threshold to 2.5% results in minor bitrate increase (<0.5 Kbps) compared to setting it by 12 which causes significant increase of bitrate (>12 Kbps).

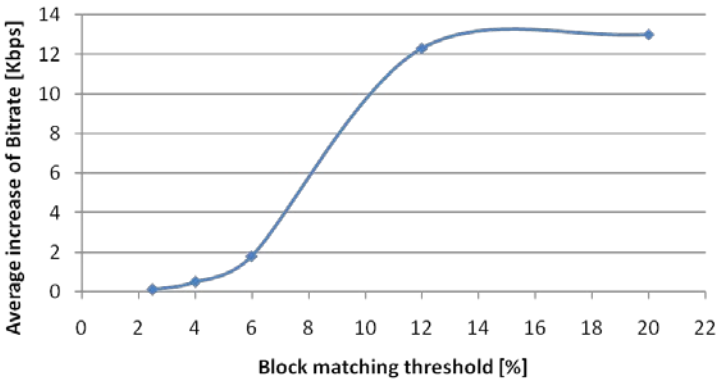
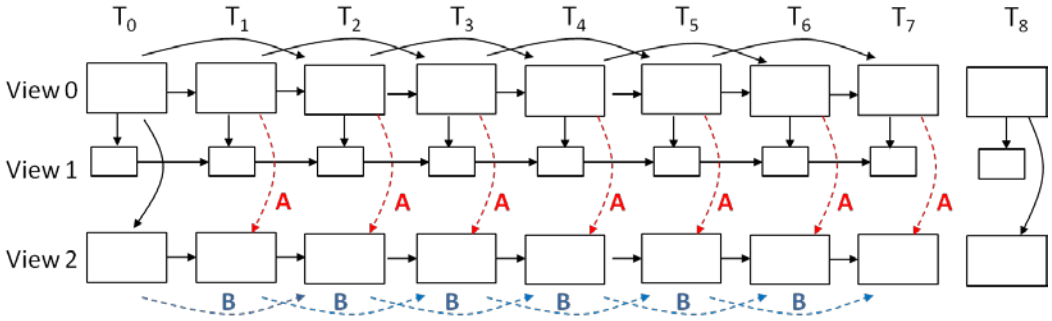


Figure 5-26 Effect of using different block matching thresholds on bitrate

5.3.4 Proposed prediction architecture

Reference frame selection and reference frame ordering are defined based on the results that are presented in subsections 5.3.2 and 5.3.3. Figure 5-27 depicts prediction architecture for mixed spatial-resolution multi-views video coding, where Group Of Picture (GOP) is eight. It deploys low spatial-resolution frames in the middle view. Dashed red and blue arrows are reference frames that are used when conditions A and B are true. When the amount of inter-view prediction blocks for low spatial-resolution frame is higher than the threshold, then condition A is true. Similarly, when temporal predicted blocks for frame belongs to base view is higher than the threshold, then condition B is true. Threshold is set to 2.5% that reflects insignificant amount of block matching.

Two recent temporal frames are used to predict frames in base view, while nearest temporal and spatial reference frames are used during LR frames prediction. For FR frames that belong to the third view, there are four reference frame selection cases. They are illustrated in Table 5-9, where REF is the reference frame. Spatial and 2nd temporal reference frames are selected when their expected amount of block matching are significant.



Condition A is true when the amount of spatial prediction using S_0 REF for LR frame ≥ 2.5% while condition B is true when the amount of temporal prediction using T_1 REF for FR frame ≥ 2.5%

Figure 5-27 Proposed prediction architecture for mixed spatial-resolution MVC

Table 5-9 Four cases for reference frame selection during coding FR frames

Condition A	Condition B	1st REF	2nd REF	3rd REF
False	**False**	T_0	n/a	n/a
True	**False**	T_0	S_0	n/a
False	**True**	T_0	T_1	n/a
True	**True**	T_0	S_0	T_1

5.3.5 Results and discussions

There are three criteria when evaluating the proposed PA among other asymmetric MVC. They are computational complexity, memory consumption and coding gain. The computational complexity is represented by the encoding time that reflects the complexity for software encoder, while memory consumption is defined through the minimum number of reference frames needed for DPB. Coding gain is represented by average quality for the coded video and the average bitrate. Two prediction architectures are used for comparison. Hierarchical B-picture; HBP is 1st PA as shown in Figure 5-28-a. It is widely used in MVC, while the extended architecture based on 3D-DMB is the 2nd PA as depicted in Figure 5-28-b. Their asymmetric codec relies on IPPP coding structure that uses three reference frames. This complies with the recommendation reported by ITU-T for Digital Multimedia Broadcasting (DMB) (Antipolis, 2005). The extended architecture adds 3rd view to 3D-DMB prediction architecture where its frames use two temporal frames and one spatial frame; similar to the corresponding frames in the 2nd view. The proposed PA, HBP and extend architecture based 3D-DMB are used to encode six videos. These architectures deploy LR frames in the middle view. GOP is set to eight frames; each GOP begins with I-frame.

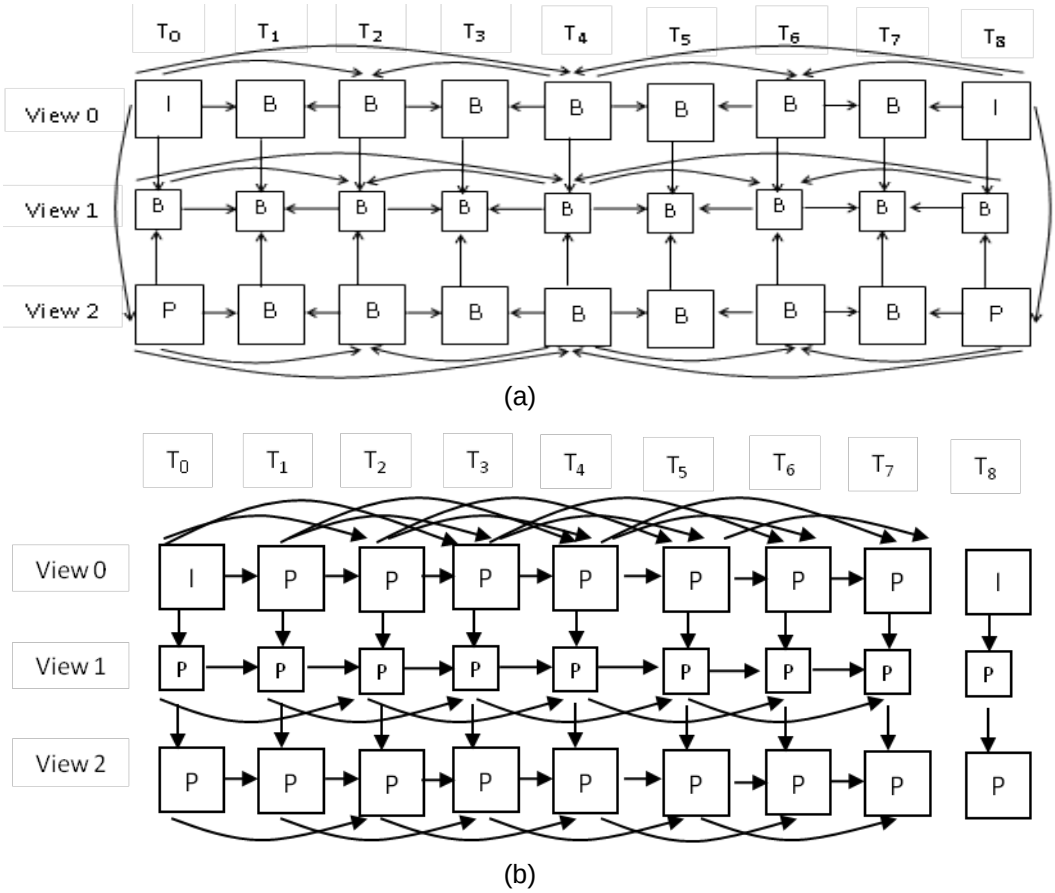


Figure 5-28 Prediction architectures a) HBP and b) Extended prediction architecture based 3D-DMB (Chen et al., 2008a; Fehn et al., 2007)

The 1st criterion is computational complexity that has two components; block matching complexity and resolution matching complexity. Block Matching (BM) finds best prediction for a given macroblock using motion and disparity estimation. The proposed prediction architecture and extended architecture based 3D-DMB use IPPP coding structure. Both architectures use three reference frames when predicting FR frames that belong to dependent view. Therefore BM checks three reference frames in forward prediction direction. HBP architecture use IBBP coding structure. Since it has three prediction directions; forward, backward and bi-prediction, BM needs to check each direction, where maximum number of reference frames is two in each direction. Therefore, BM needed for IPPP coding structure is less than IBBP structure.

Resolution matching complexity is raised during decimating or interpolating reference frame in order to match the target frame spatial-resolution. This complexity is caused at the encoder and decoder sides for disparity estimation and disparity compensation respectively. Extended prediction based 3D-DMB uses decimation and interpolation to predict LR frames and FR frames respectively. Hierarchical B-pictures architecture applies decimation for frames that belong to the base and 2nd

views in order to predict frames belonging to odd view. The proposed prediction architecture applies decimation for frames that belong to base view only. Therefore, it requires less complexity among others for resolution matching.

Memory consumption is the second criterion for evaluation, where minimum number of reference frames defines the memory size needed for prediction architecture. HBP architecture stores 10 and 14 for FR and LR reference frames as shown in Figure 5-29-a, where Blue, green and white blocks are reference frame, current frame and un-coded frame respectively. Fourteen LR frames are 4 LR (temporal prediction) plus 10 decimated FR (for IVP). Extended prediction architecture based 3D-DMB stores 8 and 6 for FR and LR reference frames respectively as shown in Figure 5-29-b. Eight FR reference frames include five FR frames and three interpolated LR frames, while six LR frames are three frames (in the middle view) and three decimated FR frames. The proposed PA stores 5 and 6 for FR and LR reference frames as depicted in Figure 5-29-c.

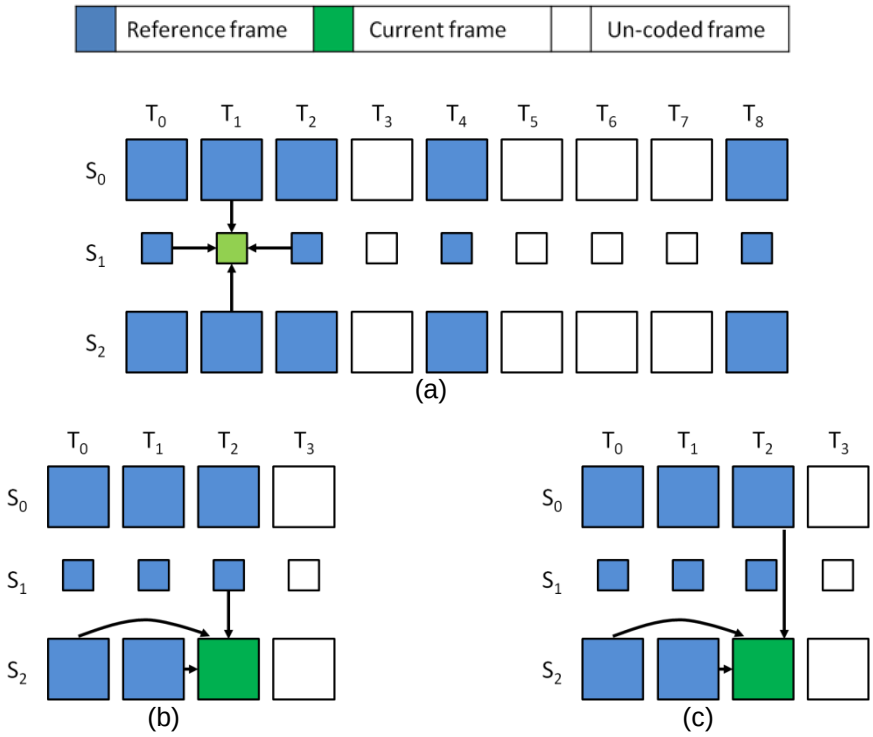


Figure 5-29 Prediction architectures for a) Hierarchical B-picture, b) Extended architecture based 3D-DMB and c) Proposed prediction architecture

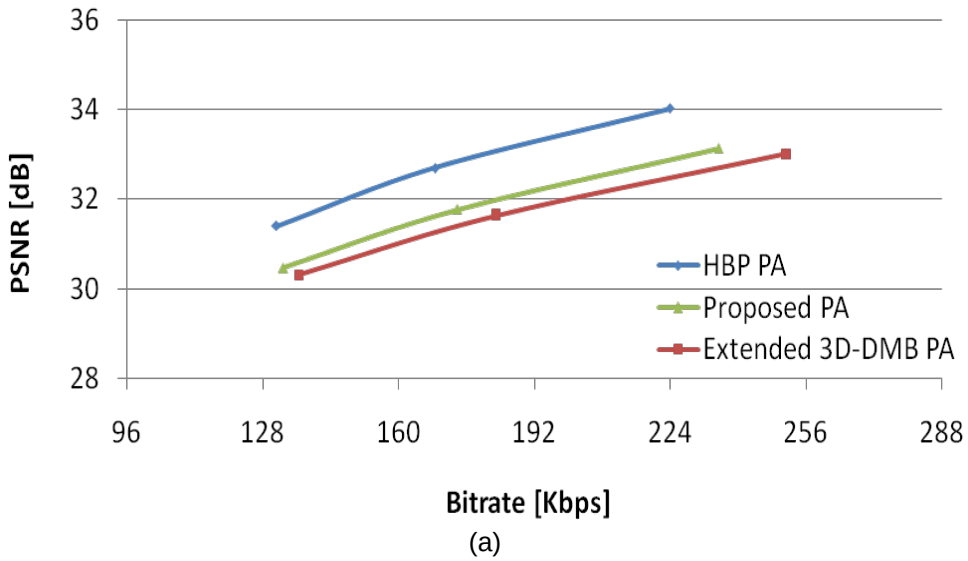
Table 5-10 shows minimum size for DPB when different prediction architectures are used. Last column shows the total amount of frames with respect to FR frames that is equal to the number of LR frames divided by 4 plus number of FR frames. The proposed prediction architecture saves significant amount of memory required for

DPB by 51.9% and 31.6% with respect to HBP and extended architecture based 3D-DMB respectively.

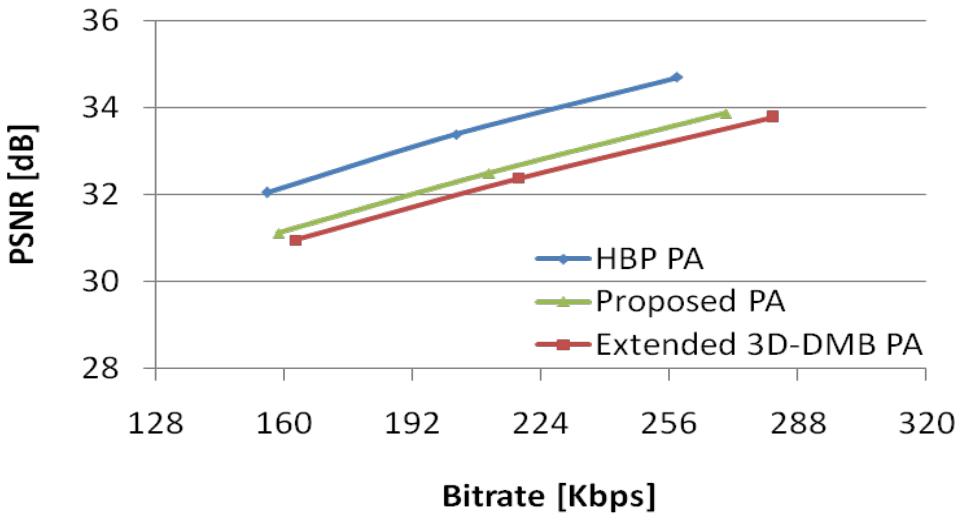
Mixed spatial-resolution multi-view videos have been coded at low bitrates using H.264/AVC based multi-view video coding. Three neighbouring views are multiplexed, where frames that belong to the middle view are decimated by a factor of two in the horizontal and vertical directions. Figure 5-30 shows rate-distortion curves for different videos, where *PSNR* (dB) and bitrate (Kbps) are presented along Y-axis and X-axis respectively. Blue, green and red curves are rate-distortion using hierarchical B-picture (HBP), proposed prediction architecture and extended architecture based 3D-DMB respectively.

Table 5-10 Minimum size for DPB for different prediction architectures

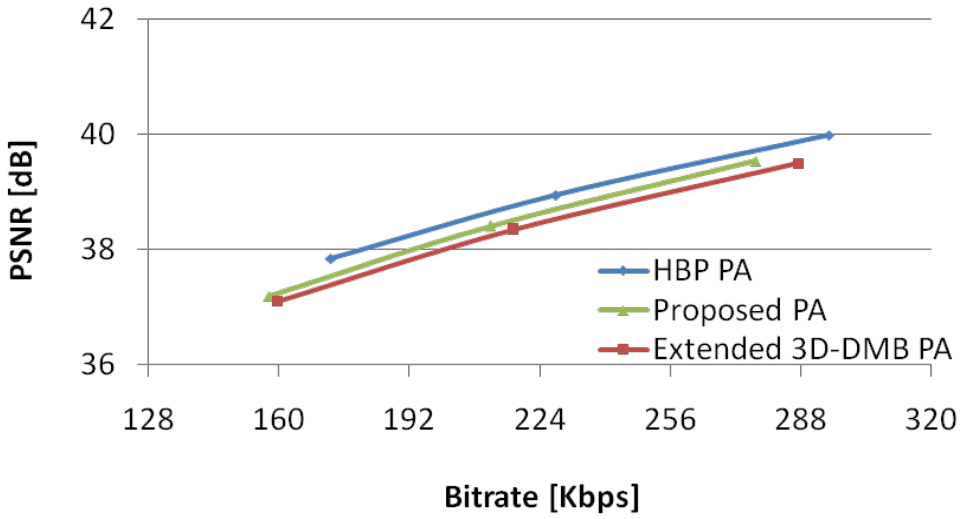
Prediction architecture	Number of FR frames	Number of LR frames	Total frames with respect to FR frames
Hierarchical B-picture	10	14	13.5
Extend architecture based 3D-DMB	8	6	9.5
Proposed architecture	5	6	6.5



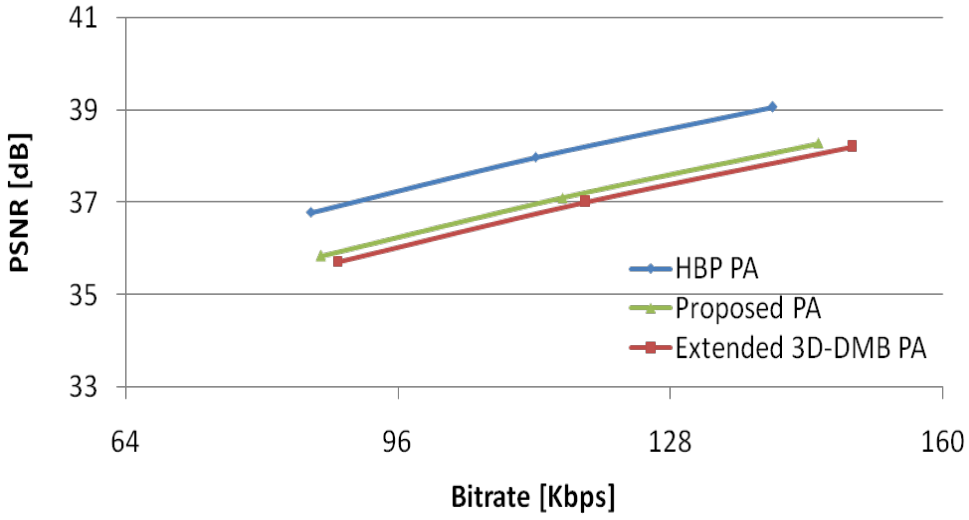
(a)


36
34
32
30
28
PSNR [dB]
128
160
192
224
256
288
320
Bitrate [Kbps]
HBP PA
Proposed PA
Extended 3D-DMB PA

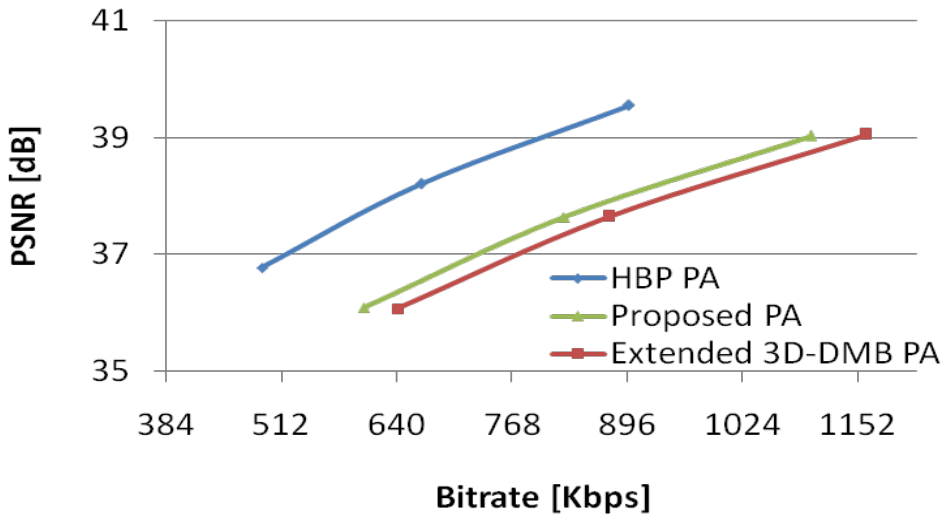
(b)


42
40
38
36
PSNR [dB]
128
160
192
224
256
288
320
Bitrate [Kbps]
HBP PA
Proposed PA
Extended 3D-DMB PA

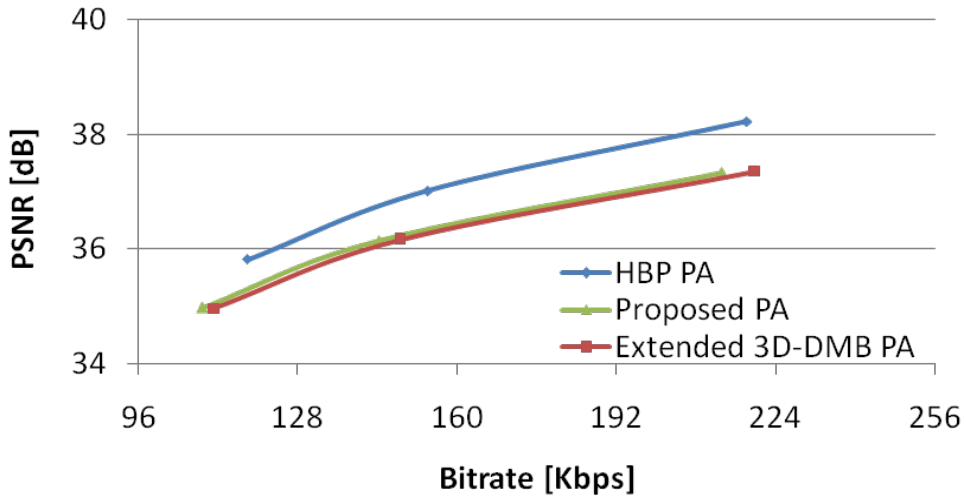
(c)


41
39
37
35
33
PSNR [dB]
64
96
128
160
Bitrate [Kbps]
HBP PA
Proposed PA
Extended 3D-DMB PA

(d)



(e)



(f)

Figure 5-30 (a-f) Rate-distortion curves for Akko & Kayo, Ballroom, Break-dancers, Exit, Race1 and Rena respectively

The proposed PA provides better coding gain than the extended architecture based 3D-DMB. The proposed PA needs less bitrate for transmitting asymmetric MVV by on average 13.1 Kbps while both obtain similar quality for decoded MVV. HBP PA gets higher coding gain than the proposed prediction architecture, where the former obtains better quality by on average 0.78 dB while requiring less bitrate by on average 24.9 Kbps with respect to the proposed prediction architecture.

Computational complexity for H.264/AVC based multi-view video coding using different prediction architectures is measured through the total encoding time. It reflects computational complexity for BM and resolution matching. All the experiments were carried out on a computer with Intel i7 CPU and memory of 16 GB. Figure 5-31 shows total time needed for encoding different videos using different

prediction architectures, where Y-axis is the total time (Hour). The proposed prediction architecture accelerates encoding by on average 57% and up to 77.5% with respect to the corresponding time needed by hierarchical B-picture architecture. It speeds up encoding by on average 14% and up to 54% with respect to the extended prediction architecture based 3D-DMB. Table 5-11 shows amount of saving S_0 and T_1 reference frames during coding FR frames that belong to 3rd view. Saving amount for S_0 is high for Exit since it has large disparity and slow objects' motion on contrary to Break-dancers that have high objects' motion (T_1 saving is 0.9%). From these results, it can be seen that the proposed PA needs less memory consumption and encoding time with respect to both; extended architecture based 3D-DMB and hierarchical B-picture prediction architectures. It gives superior coding gain than the former architecture. HBP provides best coding efficiency among other architectures that are based on IPPP coding structure at the expense of the highest computational complexity and memory consumption.

Six multi-view videos with different views have been coded for validation using the proposed PA, HBP PA and the extended architecture based 3D-DMB. The results highlight that the proposed PA saves on average 11.6 Kbps compared to the extended architecture based 3D-DMB. HBP architecture saves on average 23.6 Kbps and provides 0.63 dB better quality for the decoded asymmetric multi-view videos with respect to the proposed PA. H.264/AVC based MVC that deploys the proposed PA needs less encoding time compared to corresponding codec that uses either HBP or extended architecture based 3D-DMB by on average 49.9% (up to 58%) and 5.8% (up to 20.5%) respectively.

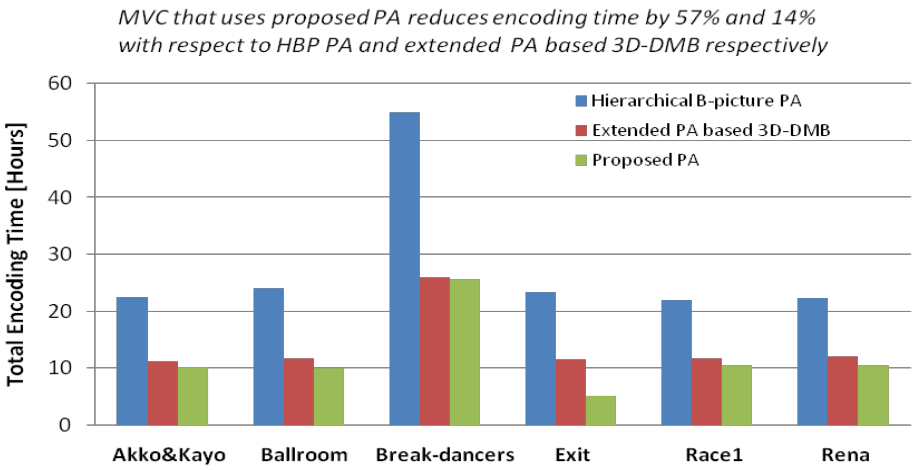


Figure 5-31 Total encoding time when using different prediction architectures

Table 5-11 the amount of saving percent for S_0 and T_1 reference frames

Dataset	Akko & Kayo	Ballroom	Break-dancers	Exit	Race1	Rena	Average
S_0 Saving %	0	1.6	0	92.9	1.6	0	16
T_1 Saving %	40.7	43.5	0.9	97.2	5.6	27.8	36

5.3.6 Proposed prediction architecture with adaptive reference frame ordering algorithm

Adaptive reference frame ordering algorithm (reported in section 4.4) has been integrated with the proposed prediction architecture. The algorithm presents efficient mechanism for reordering reference frames indices. The proposed prediction architecture contains three reference frames. The first three reference frames are indexed via '1', '010' and '011' respectively. Since the 1st frame is recent temporal frame 'T_0' which usually provides most significant role of block matching, deploying reference frame reorder algorithm would not provide bits saving for the proposed prediction architecture. However, the algorithm is beneficial when coding multi-view video that contains hard scene change. Coding frame that belong to new scene would change RFO, where the most significant reference frame becomes the nearest spatial frame instead of the recent temporal reference frame. Therefore, deploying the algorithm would be essential when coding mixed spatial-resolution multi-view video that has multiple scenes. When the first frame is coded that belongs to a new scene, the majority of coded blocks are intra-predicted. The reference frame indices will be then reordered in a way that places nearest spatial reference frame first in List 0. The new RFO is applied for the frames that belong to the neighbouring views.

Three-view video with hard scene change is generated in the context of mixed spatial-resolution MVV through multiplexing frames that belong to Akko & Kayo, Ballroom, Exit, Race1 and Rena. The sequence starts with first nine frames from Akko & Kayo, following by six frames from each MVV. Frames that belong to the middle frame are decimated while frames belonging to the surrounding views have full spatial-resolution. This sequence is coded by H.264/AVC based multi-view video coding using three prediction architectures; the proposed architecture with Adaptive Reference Frame Order (ARFO) algorithm, extended architecture based 3D-DMB and HBP prediction architectures.

The result in terms of rate-distortion is presented in Figure 5-32, where blue, red and green curves are rate-distortion when coding the video using HBP architecture, extended architecture based 3D-DMB and the proposed PA respectively. Figure 5-

33 shows amount of bits per frame when the video is coded by these prediction architectures. HBP architecture reduces the amount of bit saving with respect to the proposed prediction architecture as depicted in Figure 5-33-a. This is due to backward prediction, where the reference frame and the current frame belong to the new scene. Figure 5-33-b shows clearly bit saving with respect to extended architecture based 3D-DMB. Figure 5-34 depicts encoding time when coding MVV that has multiple scene changes. Blue, red and green bars represent total encoding time needed by HBP architecture, extended architecture based 3D-DMB and the proposed PA respectively.

The proposed architecture with adaptive reference frame ordering algorithm saves on average 28.7 Kbps and 35.4 Kbps with respect to HBP architecture and the extended architecture based 3D-DMB respectively. It provides similar quality for decoded asymmetric multi-view video to the corresponding video coded via extended architecture based 3D-DMB. HBP provides better quality by on average 0.38 dB compared to the corresponding video that is coded by the proposed prediction architecture. The proposed prediction architecture accelerates encoding time by on average 64% and 33% with respect to the corresponding time needed by hierarchical B-picture architecture and the extended PA based 3D-DMB.

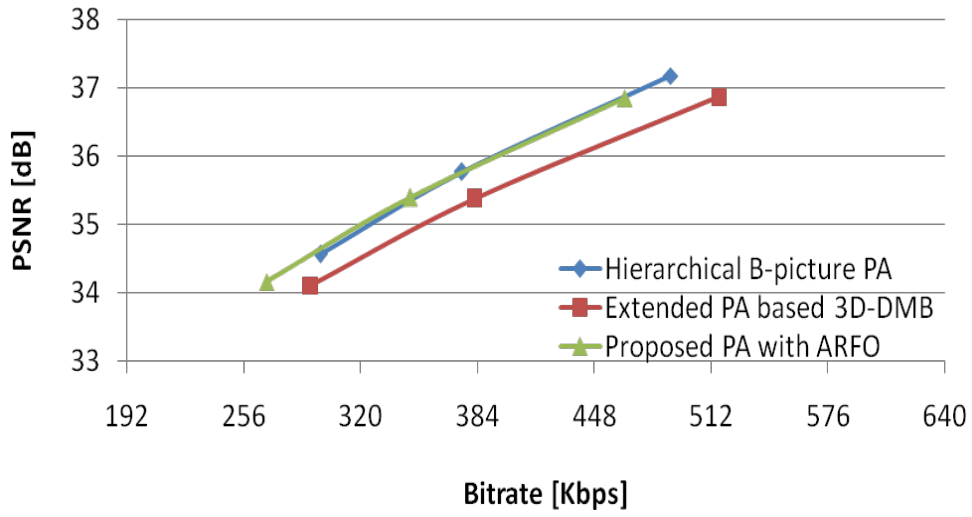
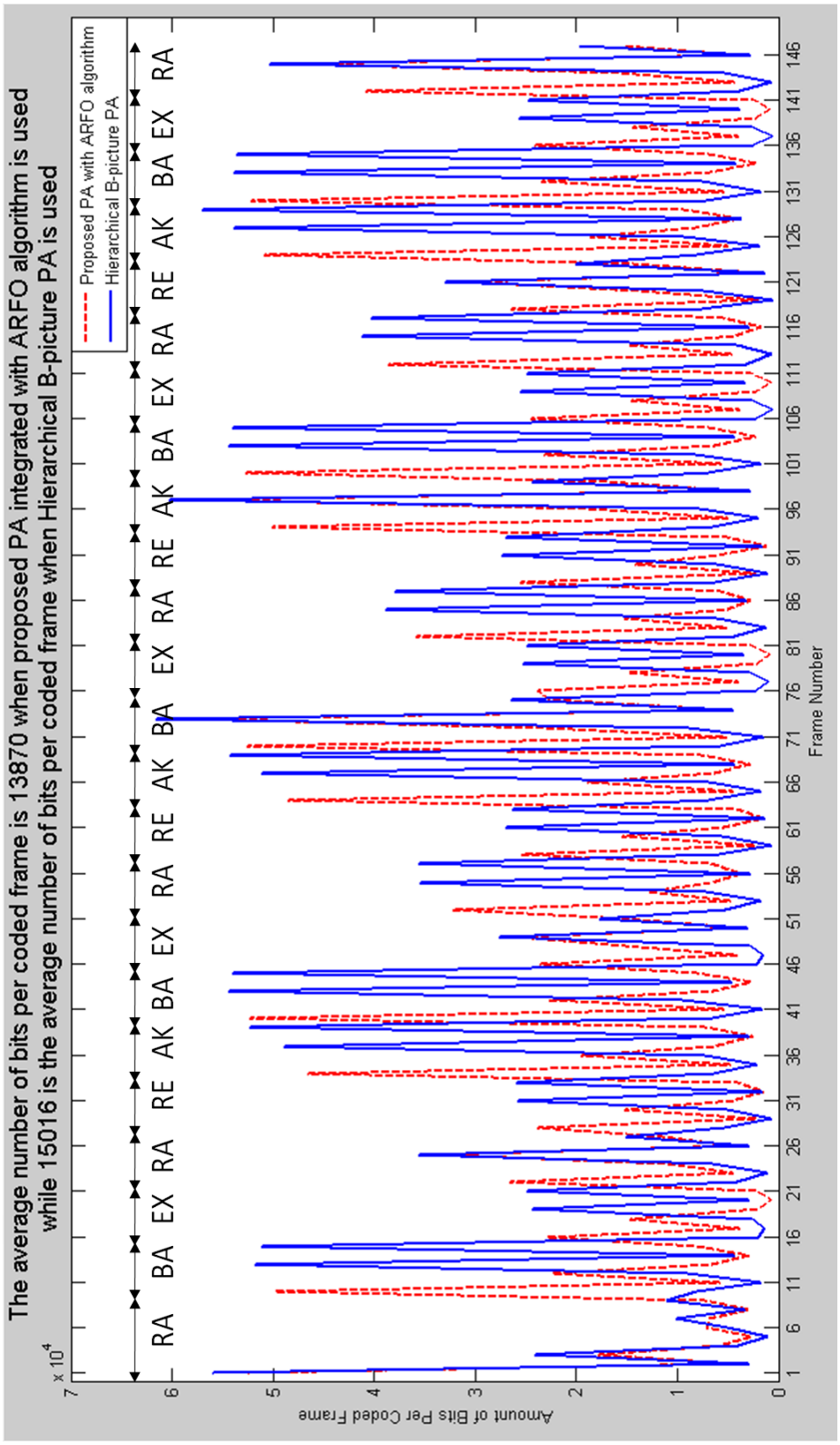
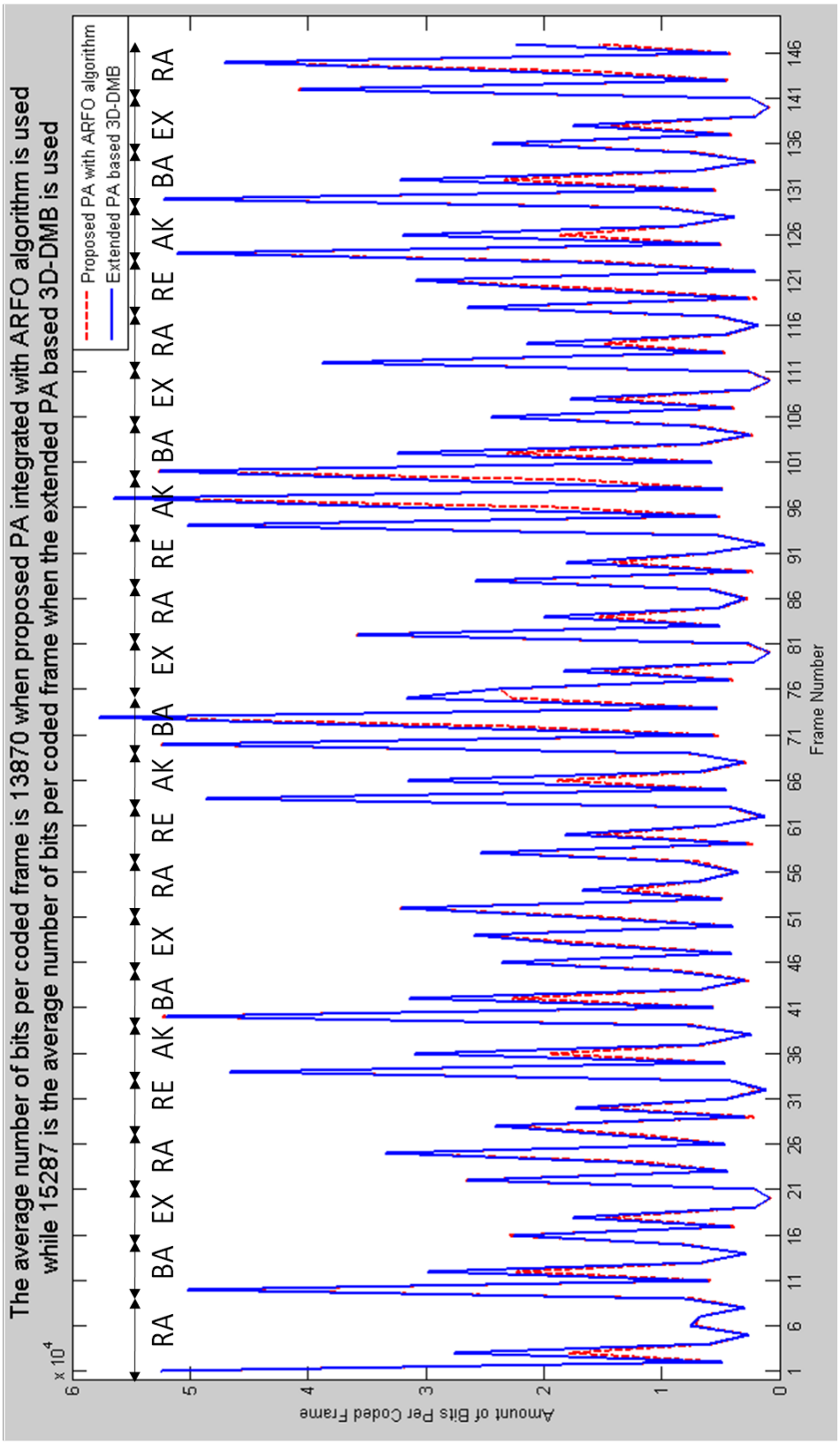


Figure 5-32 Rate-distortion curves when coding MVV that contains hard scene change


The average number of bits per coded frame is 13870 when proposed PA integrated with ARFO algorithm is used
while 15016 is the average number of bits per coded frame when Hierarchical B-picture PA is used
Proposed PA with ARFO algorithm
Hierarchical B-picture PA
Amount of Bits Per Coded Frame
Frame Number
x 10^4
RA BA EX RA RE AK BA EX RA RE AK BA EX RA RE AK BA EX RA RE AK BA EX RA

(a)



(b)

Figure 5-33 Amount of bits per frame using proposed prediction architecture with HBP PA and b) Extended PA based 3D-DMB a)

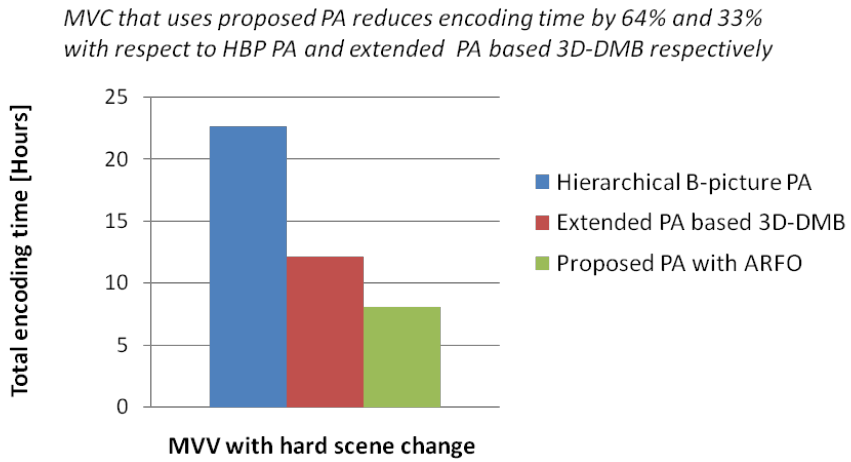


Figure 5-34 Total encoding time when coding MVV with hard scene changes using different prediction architectures

5.3.7 Conclusions

Prediction architecture is proposed based on the block matching statistics for full and low spatial-resolution frames. Nearest temporal and spatial reference frames are selected during coding LR frames. Full spatial-resolution frames that belong to dependent view use two temporal frames and FR spatial reference frames. For full spatial-resolution frames, temporal and spatial reference frames are dynamically skipped when their expected amount of block matching are insignificant. The proposed prediction architecture is compared to the extended architecture based 3D-DMB and hierarchical B-picture prediction architectures in terms of computational complexity, memory consumption and rate-distortion. From the results, the proposed prediction architecture saves significant amount of memory required for DPB by 51.9% and 31.6% with respect to HBP and extended architecture based 3D-DMB respectively.

The proposed prediction architecture accelerates encoding by on average 57% and up to 77.5% with respect to the corresponding time needed by hierarchical B-picture architecture. It speeds up encoding by on average 14% and up to 54% with respect to extended prediction architecture based 3D-DMB. The proposed prediction architecture needs less bitrate for coding asymmetric multi-view video than extended PA based 3D-DMB by on average 13.1 Kbps while both obtain similar quality for decoded multi-view video. The MVC that uses HBP PA has higher coding performance than the corresponding codec that uses the proposed PA. HBP PA provides higher quality for coded videos by on average 0.78 dB while achieves less bitrate by on average 24.9 Kbps with respect to the proposed PA.

Adaptive reference frame ordering algorithm is integrated with the proposed prediction architecture to provide efficient mechanism for coding multi-view video that contains several scenes. When scene changes the adaptive reference frame ordering algorithm modifies spatial reference index to be the first in the List 0. Therefore, next frames located at the same time slice will be predicted via spatial reference frame that needs single bit for indexing. The proposed architecture with reference frame reorder algorithm saves on average 28.7 Kbps and 35.4 Kbps with respect to HBP architecture and extended architecture based 3D-DMB respectively. It provides similar quality for decoded asymmetric MVV to the corresponding MVV coded via extended architecture based 3D-DMB. HBP provides better quality by on average 0.38 dB compared to the corresponding video that is coded by the proposed prediction architecture. The proposed PA accelerates encoding time by on average 64% and 33% with respect to the corresponding time needed by hierarchical B-picture architecture and the extended PA based 3D-DMB.

Although suppression theory provides acceptable justification for deploying mixed spatial-resolution frames, this type of coded video causes eye fatigue when it is watched for several minutes (Jain et al., 2014). On the other hand, it could be used in free-viewpoint video (Garcia et al., 2010a). Since the interpolated frames suffer from blurriness, the next section will investigate visual quality enhancement for the interpolated frames using embedded information in neighbouring full spatial-resolution frames.

5.4 Visual quality enhancement algorithm for interpolated frames

5.4.1 Introduction

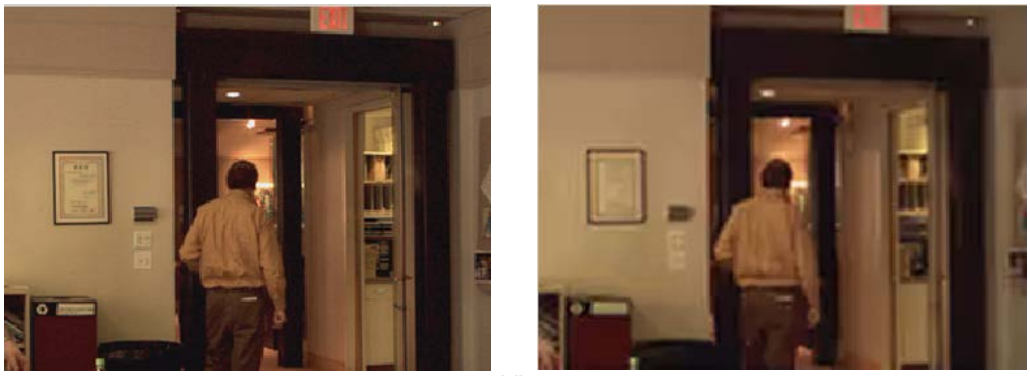
This section focuses on enhancing visual quality for the interpolated frames. The coded LR frames at low bitrates suffer from blockiness and blurriness artefacts when they are interpolated at the receiver side. To realise both artefacts, H.264/AVC based stereoscopic video coding is used to encode mixed spatial-resolution videos at low bitrates. The first two views from Akko & Kayo, Ballroom, Break-dancers, Exit, Race1 and Rena have been coded. Frames that belong to the dependent view are decimated prior to compression while the decoded LR frames have been interpolated before display. Figure 5-35 shows un-coded and interpolated frames for the first frame that belong to the dependent view. The first column presents un-coded frame while the second column shows coded LR frame after interpolation. From this figure,

the details that are visible in un-coded frame are degraded significantly with respect to the interpolated frame. Coding LR frames at low bitrates causes blockiness artefacts while interpolating it prior to display, reduce the details significantly. Therefore, the blockiness artefacts are magnified by the negative effect from interpolation. This entails reducing the visual quality of interpolated frames significantly with respect to un-coded frames. Since there are two sets of coded frames; FR and LR frames, the information that exists in FR frames are exploited to enhance visual quality for the interpolated frames.

(a)

(b)

(c)



(d)

(e)

(f)

Figure 5-35 (a-f) Un-coded frame versus interpolated frame for Akko & Kayo, Ballroom, Break-dancers, Exit, Race1 and Rena respectively

In the next subsection, the relationship among residual error for inter-view predicted blocks and disparity compensation will be highlighted. The proposed visual enhancement algorithm will be presented that is based on disparity compensation, where its applications, display and inter-view prediction are discussed. The integration among the proposed visual enhancement algorithm and the proposed prediction architecture is then presented.

5.4.2 Residual error for disparity compensation

This subsection investigates the relationship between residual error for inter-view predicted block and disparity compensation. The frames that belong to base view are FR while LR frames are used in dependent view. During inter-view prediction, reference frame is decimated to match spatial-resolution for the current frame that entails storing two copies of reference frame; FR and LR frames. Block matching uses blocks that belong to LR reference frame to predict blocks that belong to LR target frame. Residual energy is generated through subtracting reference (Predicted) block from the target (Original) block as shown in following equation.

$$Error\ Signal = Original\ Signal - Predicted\ Signal \tag{5-2}$$

The previous equation is applicable when coding FR frame. The next equations show block matching when it is deployed for LR and FR frames.

$$BM\ Error_{\ LR} = F2_{LR} - F1_{LR} \tag{5-3}$$

$$BM\ Error_{\ FR} = F2_{FR} - F1_{FR} \tag{5-4}$$

In equations 5-3 and 5-4, both *F1* and, *F2* refer to reference (predicted) and target (original) blocks respectively. When $BM\ Error_{FR}$ is available, then the target block would be computed directly from eq. 5-4. In mixed spatial-resolution MVC, $BM\ Error_{FR}$ is not available; however it could be estimated by interpolating $BM\ Error_{LR}$. This error signal is available during disparity compensation. Figure 5-36 illustrates how to estimate the target block ($F2'_{FR}$) using information available from the reference block ($F1_{FR}$) that belongs to FR frame during disparity compensation. The information includes disparity vector that is obtained by disparity estimation and residual signal ($BM\ Error_{LR}$), where both are available at the encoder and decoder sides.

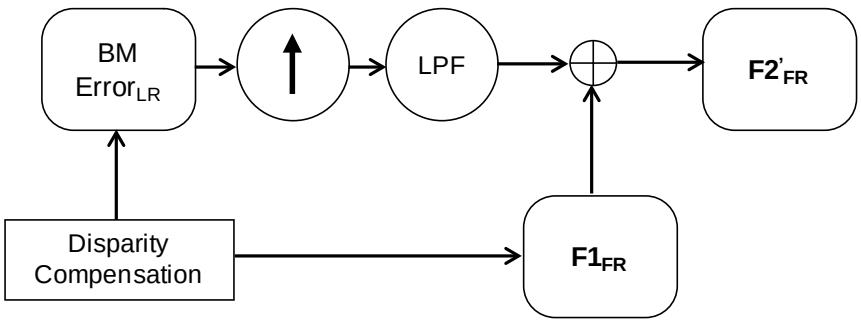


Figure 5-36 Estimating FR frame using disparity compensation

Utilising the information from disparity compensation has been studied previously via *Yang et al.* (Yang et al., 2009) that is reported in subsection 3.3.1.1.2. Their study aims to reduce the amount of time needed for interpolation in order to decrease the amount of decoding complexity at the receiver side. In this subsection, all inter-view predicted blocks are used to estimate the corresponding FR blocks during interpolating coded LR frame that belongs to the dependent view.

There are three types of blocks for a coded LR frame, they are intra, temporal and inter-view predicted blocks. For inter-view predicted blocks, there are two types of blocks. First types of blocks are predicted, where their energy is zero while second type, the predicted blocks are associated with their residual. First type of blocks ($F2'_{FR}$) are copied directly from blocks that belong to the FR frame ($F1_{FR}$) as shown in eq. 5-5. The second type of blocks ($F2'_{FR}$) are estimated by adding corresponding prediction from $F1_{FR}$ to interpolated signal from BM $Error_{LR}$ as shown in eq. 5-6. For both inter-view predicted blocks, disparity compensation uses samples at integer and sub-pixel positions to obtain $F1_{FR}$.

$$F2'_{FR} = F1_{FR} \tag{5-5}$$

$$F2'_{FR} = F1_{FR} + BM\ Error'_{FR} \tag{5-6}$$

The key point in enhancing visual quality of interpolated frame is the amount of correlation among estimated residual (eq. 5-6) and actual residual signals. To explore this correlation, two different experiments have been carried out. The first experiment uses symmetric FR stereoscopic video coding, where the first two reference frames and their following frames that belong to dependent view are exploited and stored separately as depicted in Figure 5-37 (F_1 and F_3 that are located at the time slices T_0 and T_1). The second experiment uses mixed spatial-resolution stereoscopic video coding, where the same coding setup (quantisation parameter) is applied and the corresponding frames are extracted.

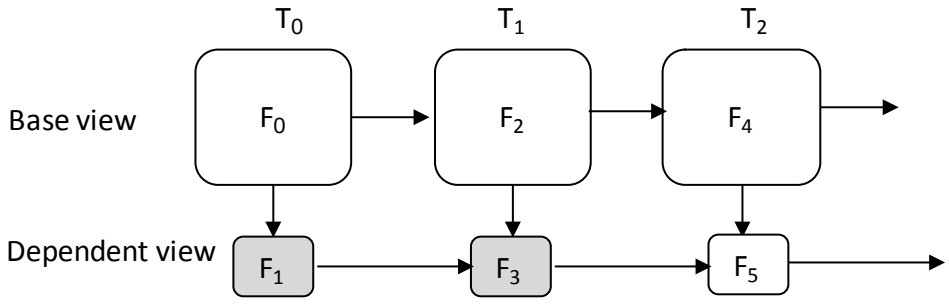


Figure 5-37 Low spatial-resolution frames that are used to compute residual correlation among actual and estimated signals

An analysis is carried out to measure the correlation among estimated and actual residual energies. Actual energy is obtained from first experiment by subtracting the predicted from original blocks. In the second experiment, when *BM Error*$_{LR}$ exists during block matching, then this signal is interpolated to obtain estimated residual signal. The correlation between actual and estimated residual signals is measured. Tables 5-12 and 5-13 present inter-view prediction analysis for F_1 and F_3 respectively. During coding, F_1 frame uses inter-view prediction in the majority of blocks while F_3 frame relies mainly on temporal prediction. Total amounts for inter-view predicted blocks are on average 81% and 17% while 37% and 2.2% are the amounts of inter-view predicted blocks that do not have residual signal for F_1 and F_3 respectively. The amounts of inter-view predicted blocks that are associated with the residual signal are 44% and 14.8% for F_1 and F_3. The average correlation for actual and estimated residual signals is presented in the last column that measures the correlation among inter-view predicted blocks that are associated with the residual signals.

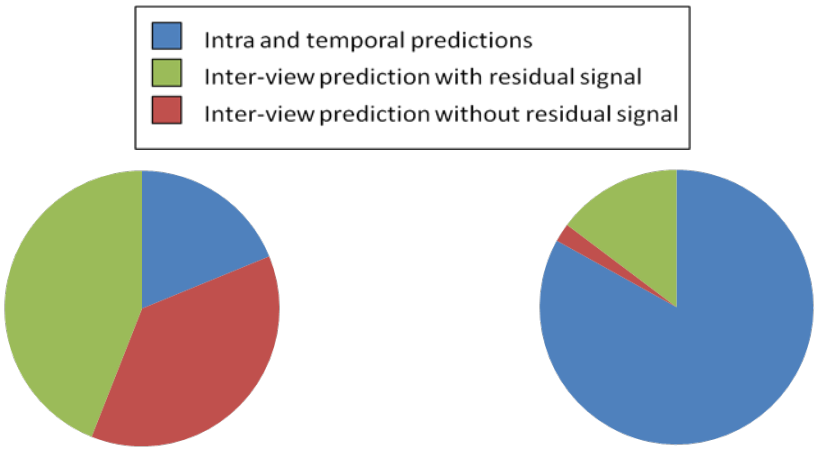
The net results for F_1 and F_3 frames that belong to the dependent view are summarised as depicted in Figure 5-38. Blue, red and green colours reflect amount of intra plus temporal, inter-view prediction without residual and inter-view prediction that is associated with residual signal respectively. Less than half the amounts of target blocks need to interpolate residual signal for frames that follow key frames (e.g. F_1). For frames that follow non-key frames (e.g. F_3), the total amount of inter-view prediction is significantly less than the amount of intra and temporal predicted blocks. Figure 5-39 depicts the amount of error correlation (per 8×8 block) for F_1 frame, where X-axis and Y-axis are number of blocks and correlation value for residual signal among actual and estimated signals.

Table 5-12 Statistical analysis of inter-view prediction for F_1 frame

Dataset	Intra predicted blocks %	IVP %	IVP % where Error = 0	IVP % where Error ≠ 0	Average correlation error
Akko & Kayo	8.67	91.33	58	33.33	0.2
Ball	18	82	30	52	0.24
Break-dancers	27.6	72.4	8.98	63.41	0.27
Exit	17.67	82.33	30.67	51.67	0.29
Race1	26.67	73.33	28.33	45	0.27
Rena	14.33	85.67	67	18.67	0.22
Average	**18.82**	**81.18**	**37.16**	**44.01**	**0.25**

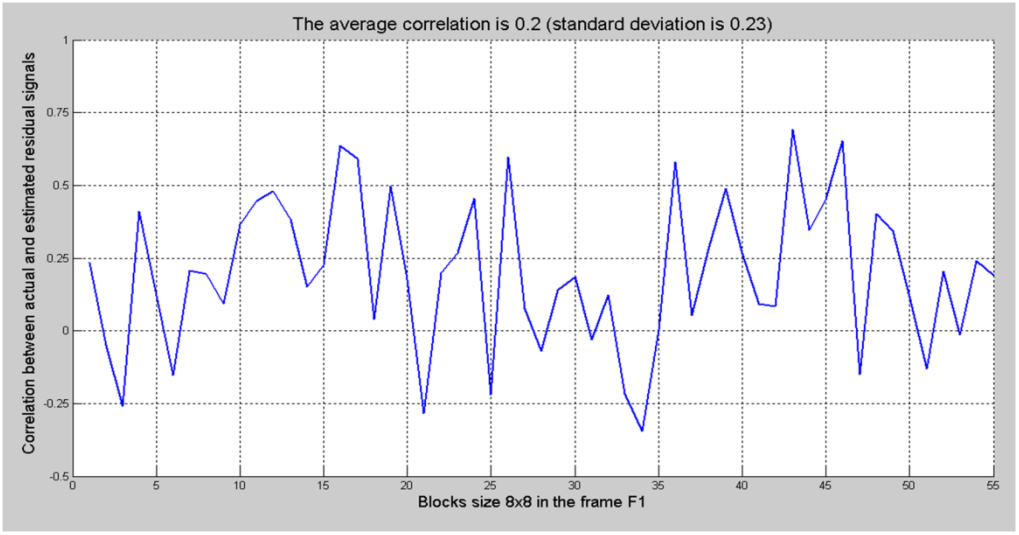
Table 5-13 Statistical analysis of inter-view prediction for F_3 frame

Dataset	Intra / Temporal predicted blocks %	IVP %	IVP % where Error = 0	IVP % where Error ≠ 0	Average correlation error
Akko & Kayo	86.33	13.67	2.33	11.33	0.14
Ball	89.33	10.67	1	9.67	0.29
Break-dancers	73.57	26.43	2.6	23.83	0.24
Exit[34]	99.33	0.67	0	0.67	-0.11
Race1	89.67	10.33	0.67	9.67	0.21
Rena	76.33	23.67	4.33	19.33	0.23
Average	**83.05**	**16.95**	**2.19**	**14.77**	**0.22**



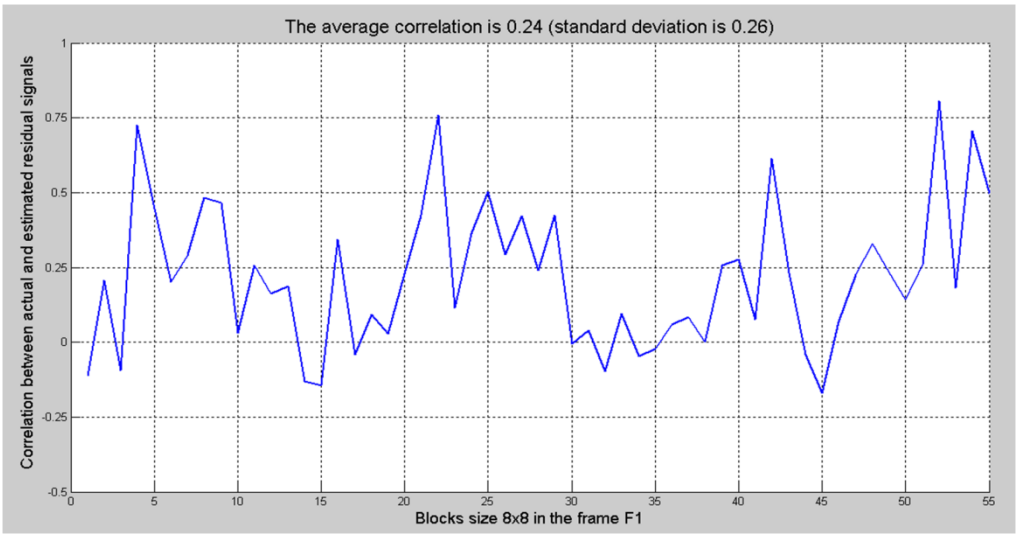
(a) (b)

Figure 5-38 Inter-view prediction statistics for dependent frame that follows a) Key frames and b) Non-key frames

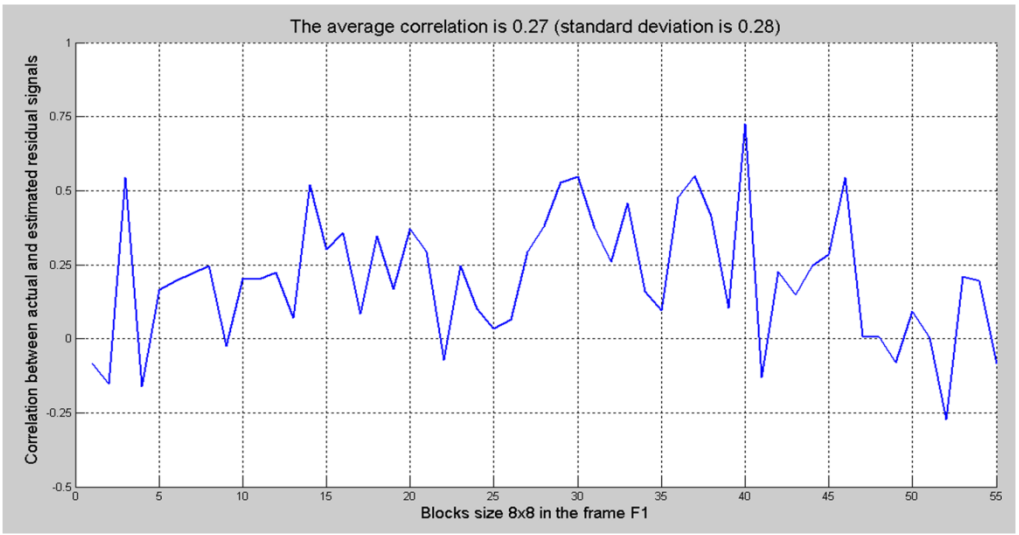


(a)

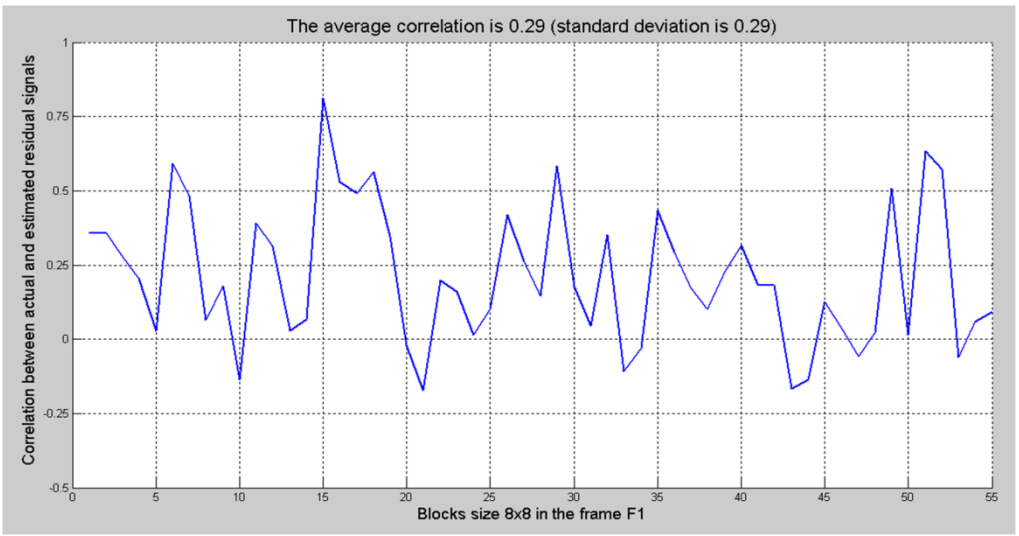
[34] It has been excluded in computing error correlation since number of IVP blocks equals two


The average correlation is 0.24 (standard deviation is 0.26)
Correlation between actual and estimated residual signals
Blocks size 8x8 in the frame F1

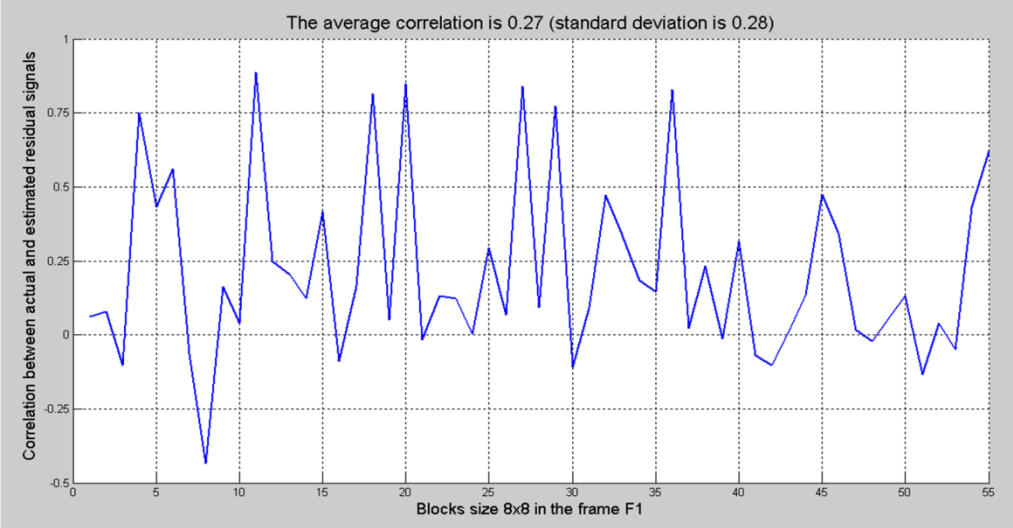
(b)


The average correlation is 0.27 (standard deviation is 0.28)
Correlation between actual and estimated residual signals
Blocks size 8x8 in the frame F1

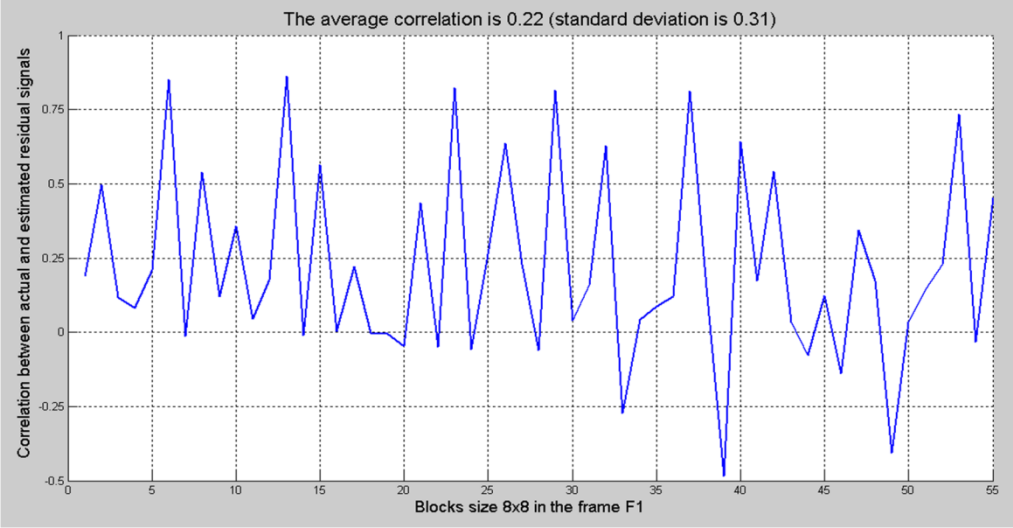
(c)


The average correlation is 0.29 (standard deviation is 0.29)
Correlation between actual and estimated residual signals
Blocks size 8x8 in the frame F1

(d)



(e)



(f)

Figure 5-39 (a-f) Residual correlation per 8×8 block among actual and estimated residual signals for Akko & Kayo, Ballroom, Break-dancers, Exit, Race1 and Rena respectively

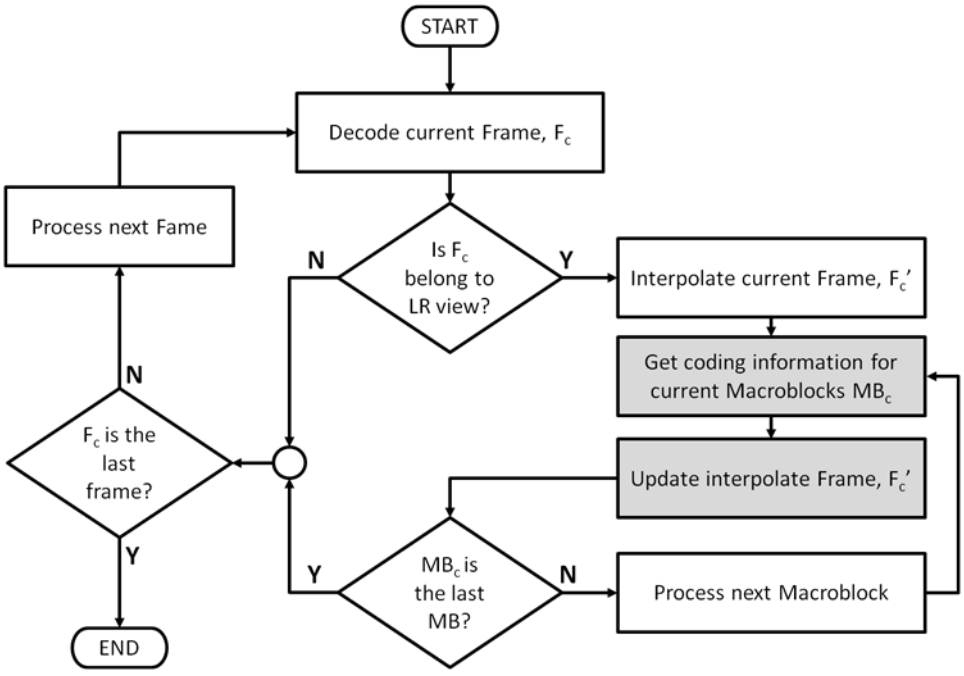
The average correlations for residual signals among actual and estimated signals are 0.25 and 0.22 for F_1 and F_3 respectively. Although the correlation is weak[35] positive, majority of the inter-view predicted blocks have 82% and 79% positive correlation for F_1 and F_3 respectively. This means that significant amount of blocks can be estimated by adding the interpolated residual signal (*BM Error'*$_{FR}$) to FR reference block (*F1*$_{FR}$).

[35] Weak correlation is based on the interpretation described by Evans (Evans, 1996)

5.4.3 Proposed visual enhancement algorithm

The proposed Visual Enhancement (VE) algorithm is depicted in Figure 5-40-a where the main steps are shaded by grey colour. Figures 5-40-b and 5-40-c show steps for coding information utilisation and frame updating procedures respectively.

The proposed algorithm starts when LR frame is decoded and interpolated as shown in Figure 5-40-a. The proposed algorithm performs two steps; utilising coding information and updating the interpolated frame. The first step searches for inter-view predicted blocks, where the residual signal and disparity vector are extracted. Disparity vector is scaled by a factor of two in the horizontal and vertical directions. Sub-pixel sample is identified from the disparity vector that will be used among scaled disparity vector to extract predicted blocks that belong to FR reference frame ($F1_{FR}$). The second step interpolates the residual signal (if it exists) by AVC interpolation filter (6-taps), then the estimated block ($F2'_{FR}$) in Figure 5-36 is computed by adding the predicted block to the interpolated residual signal (*BM Error'*$_{FR}$). This estimated block will replace the corresponding block that exists in the interpolated coded frame. The algorithm repeats these steps for all inter-view predicted blocks prior to saving the output frame.



(a)

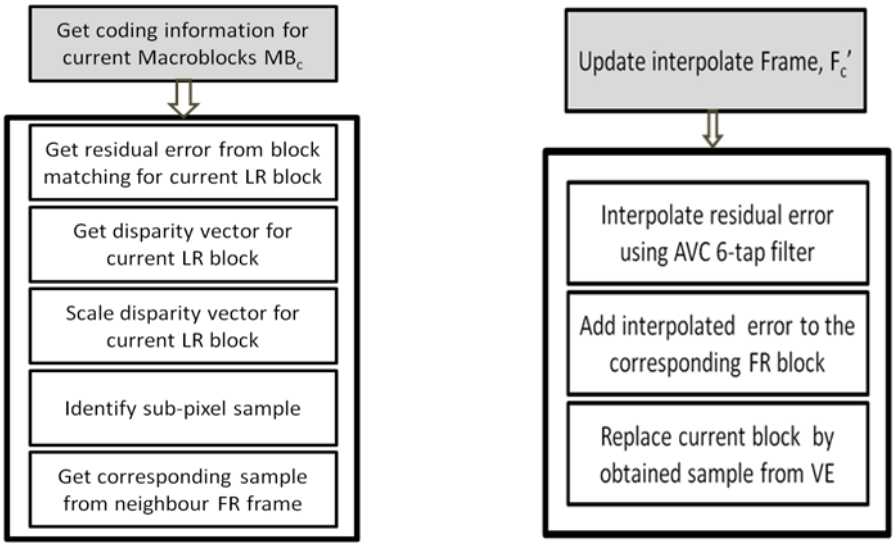


Figure 5-40 Proposed visual enhancement algorithm: a) Main algorithm, b) Coding information utilisation and c) Frame update procedure

The following Figures (5-41 to 5-46) show examples when VE algorithm is deployed for Akko & Kayo, Ballroom, Break-dancers, Exit, Race1 and Rena multi-view videos respectively. These Figures present luminance component for un-coded, interpolated coded fame using AVC filter and visually enhanced frame that is obtained by VE algorithm, where conventional decimation method is applied for FR reference frame prior to inter-view prediction.

(a)

(b)

(c)

Figure 5-41 (a-c) VE example using Akko & Kayo, where a) un-coded, b) interpolated via AVC filter and c) visual enhanced frame

(a)

(b)

(c)

Figure 5-42 (a-c) VE example using Ballroom, where a) un-coded, b) interpolated via AVC filter and c) visual enhanced frame

(a)

(b)

(c)

Figure 5-43 (a-c) VE example using Break-dancers, where a) un-coded, b) interpolated via AVC filter and c) visual enhanced frame

(a)

(b)

(c)

Figure 5-44 (a-c) VE example using Exit, where a) un-coded, b) interpolated via AVC filter and c) visual enhanced frame

(a)

(b)

(c)

Figure 5-45 (a-c) VE example using Race1, where a) un-coded, b) interpolated via AVC filter and c) visual enhanced frame

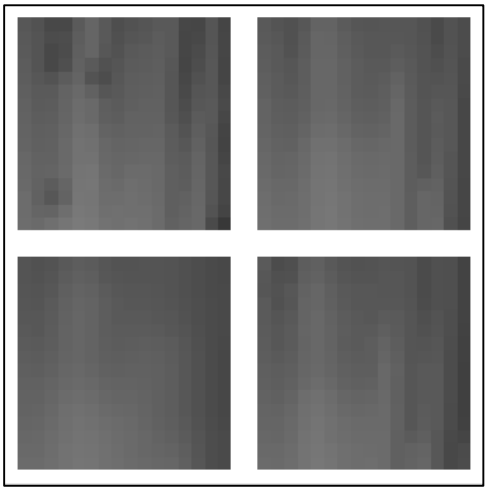
(a)

(b)

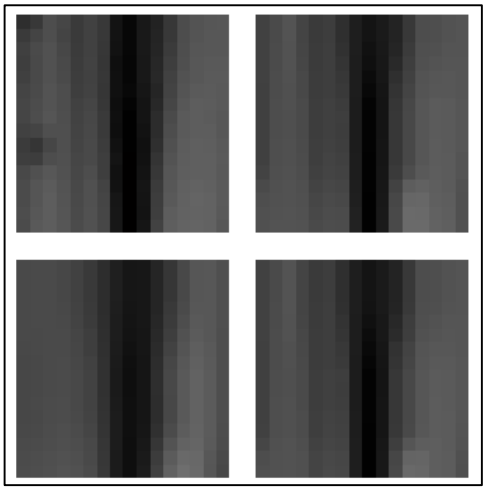
(c)

Figure 5-46 (a-c) VE example using Rena, where a) un-coded, b) interpolated via AVC filter and c) visual enhanced frame

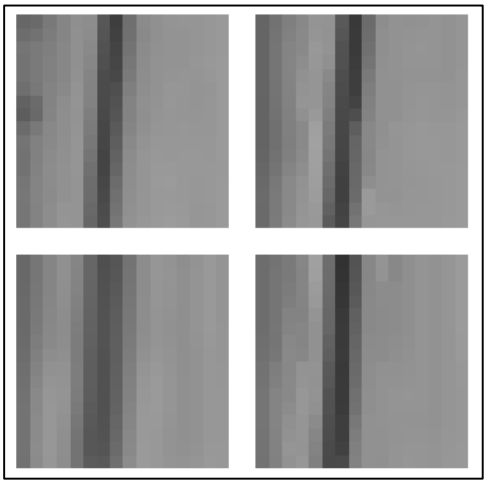
Figure 5-47 and Figure 5-48 present a set of examples for blocks size of 16 × 16 pixels (luminance component) using the 1st frame of Akko & Kayo MVV. Each figure contains four blocks, where first two blocks are un-coded and coded blocks using symmetric spatial-resolution stereoscopic video coding. The remaining two (mixed spatial-resolution stereoscopic video coding) are coded blocks that are obtained by AVC interpolation filter and VE algorithm respectively. Figure 5-47 (a-i) presents visual enhanced blocks, where the Sum Square Error (SSE) for residual signal is zero. Figure 5-48 (a-i) shows examples of twelve different blocks, where their corresponding SSE for residual signal are 13, 55, 72, 76, 93, 95, 125, 258, 4834, 5433, 7665 and 9723 respectively. It can be seen from both set of examples that the amount of blurriness that exists in visually enhanced frame or block level is less than the corresponding one that is interpolated by AVC filter. In another word, proposed VE algorithm increases frame edge's sharpness with respect to AVC interpolation filter.



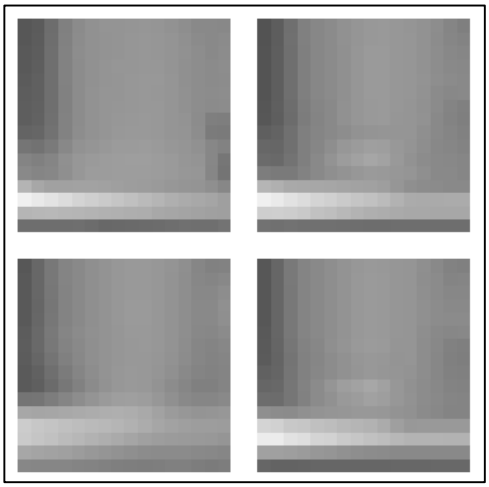
(a)



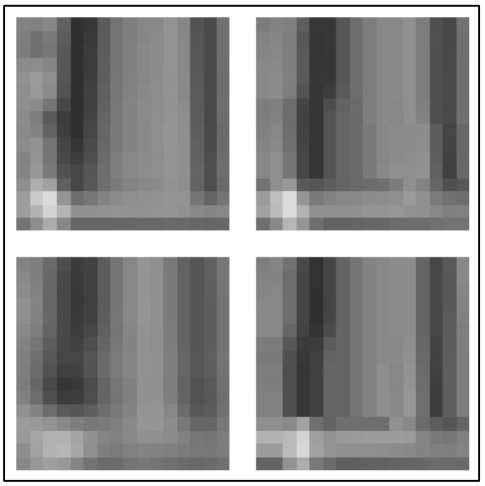
(b)



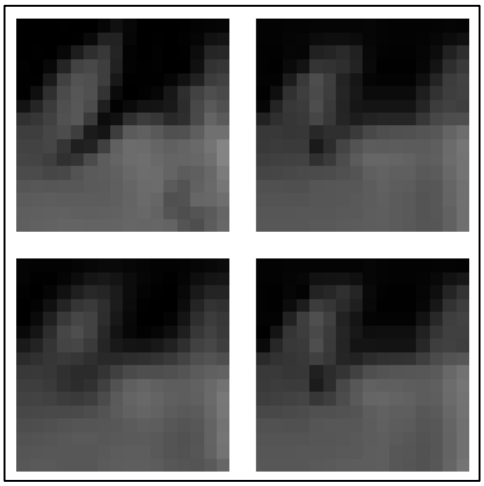
(c)



(d)



(e)



(f)

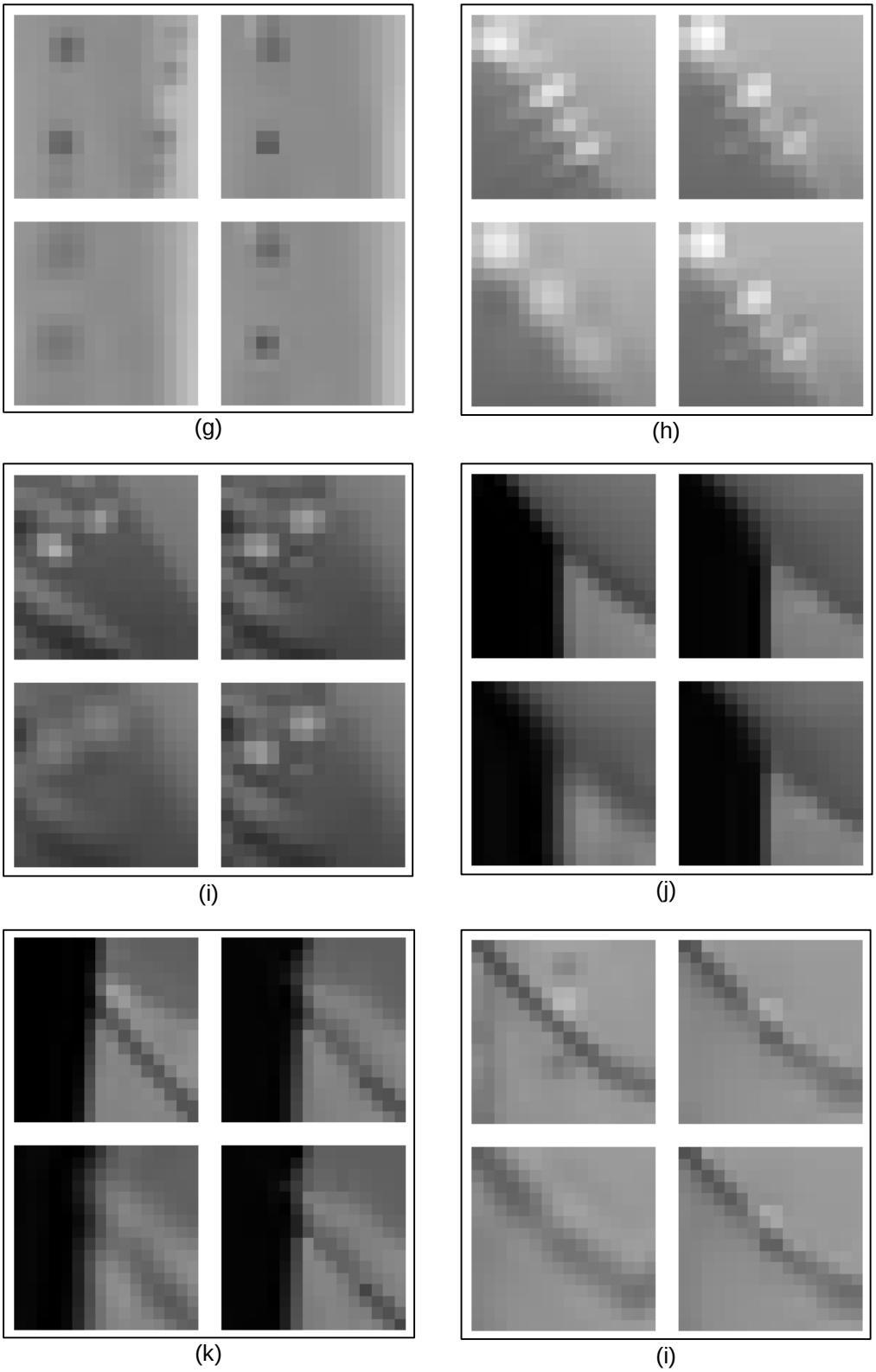
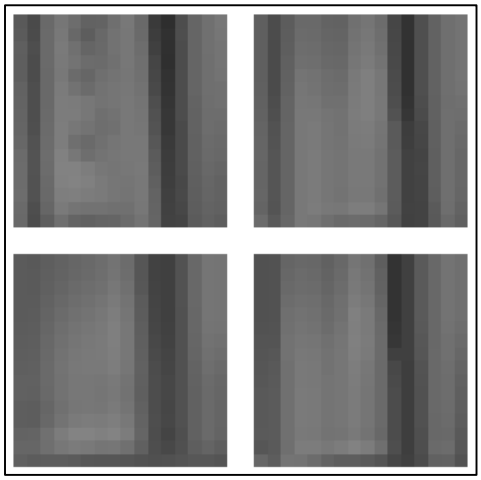
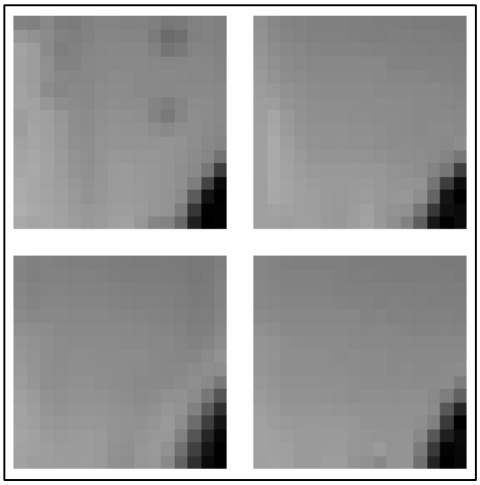


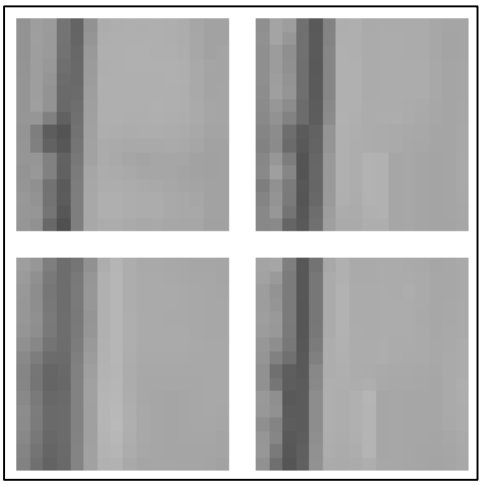
Figure 5-47 (a-i) Visual enhanced blocks, where their residual signal is zero



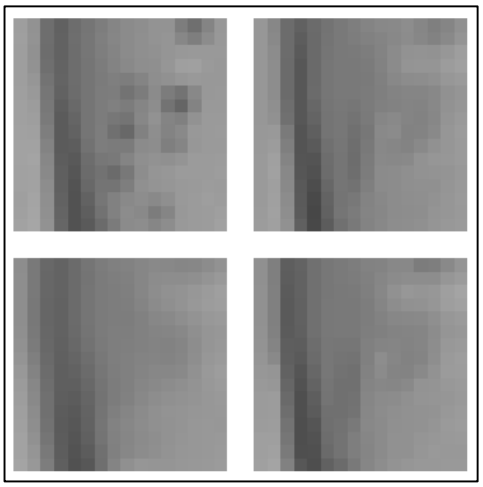
(a)



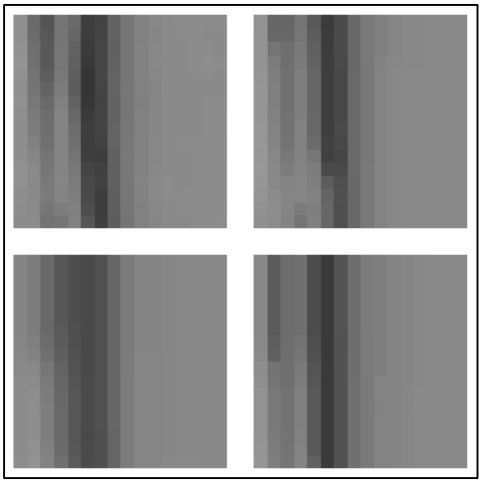
(b)



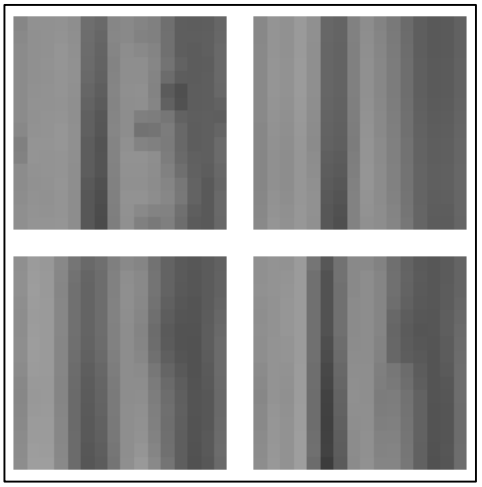
(c)



(d)



(e)



(f)

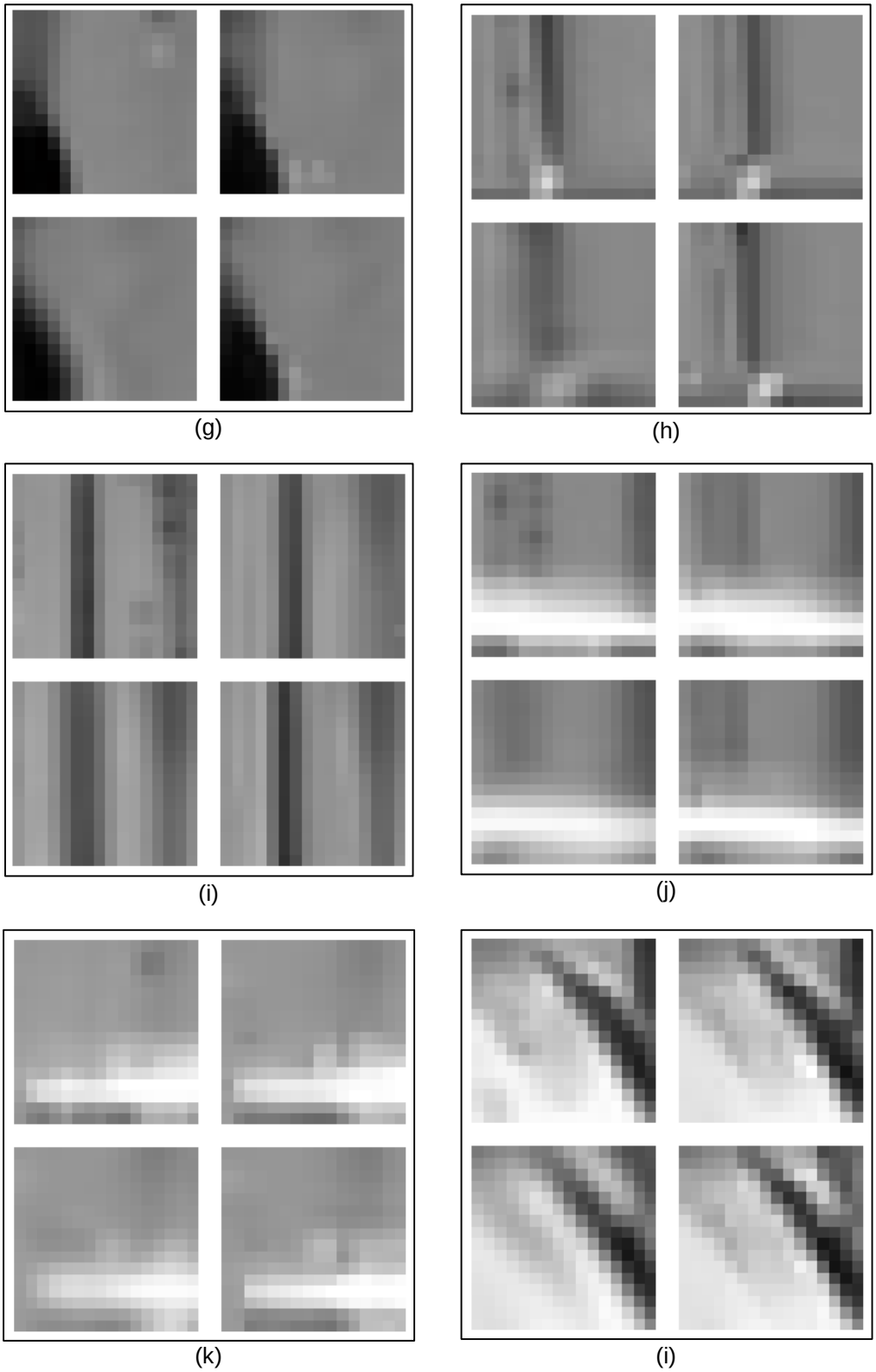


Figure 5-48 (a-i) Visual enhanced blocks that are associated with residual signal during disparity compensation

5.4.4 Proposed algorithm applications

There are two applications for the proposed visual enhancement algorithm. The first application is reducing the amount of blurriness in the interpolated frame prior to display. The second application is improving inter-view prediction when visual enhanced frames are used to predict neighbouring full spatial-resolution frames.

In context of the 1st application; display, the proposed VE algorithm is applied to the interpolated frames using mixed spatial-resolution stereoscopic video coding, where dependent view has LR frames. From subsection 5.2.2, there are two decimation methods for disparity estimation. They are conventional and high performance decimation methods. Both are used separately to evaluate the visual quality of frames that are obtained by the proposed VE algorithm. The algorithm is applied for the first two frames that belong to dependent view (F_1 and F_3 as shown in Figure 5-37). Four objective metrics are used to compare the coded LR frames that are interpolated by AVC filter and the corresponding frames that are visually enhanced by the proposed VE algorithm. These metrics are *PSNR*, *MSSIM*, *StSD* and *VQM* that is proposed by *Lee et al.*, where these metrics are outlined in section 2.3.

The first set of results using *PSNR* and *MSSIM* metrics compares visually enhanced frames when two different decimation methods are applied to FR reference frames. The first method is high performance method while the second is conventional decimation method. The following tables (5-14 to 5-19) provide the results for the first two frames; F_1 and F_3 as shown in Figure 5-37, where both frames belong to dependent view. It can be seen from these tables that the visual quality for interpolated frame is improved when VE algorithm is deployed rather than interpolating these frames by AVC filter. This is conditional when conventional decimation method is applied for FR reference frames. The *delta PSNR* improvement using over-estimated and actual measures for 1st frame (F_1) are on average 0.92 dB and 0.62 dB while 0.11 dB and, 0.09 dB are corresponding measures for the 2nd frame (F_3). When high performance method is deployed, VE algorithm provides inferior results with respect to default interpolation method.

Table 5-14 $PSNR_{actual}$ results using high performance method for F_1 frame

Multi-view video	With VE algorithm	Without VE algorithm	*Delta PSNR*
Akko & Kayo	28.34	29.42	-1.08
Ballroom	26.73	27.7	-0.97
Break-dancers	34.01	34.74	-0.73
Exit	29.56	30.77	-1.21
Race1	31.25	31.93	-0.68
Rena	32.97	33.12	-0.15
Average			**-0.8**

Table 5-15 $PSNR_{actual}$ results using conventional decimation method for F_1 frame

Multi-view video	With VE algorithm	Without VE algorithm	*Delta PSNR*
Akko & Kayo	29.47	29.16	0.3
Ballroom	28.2	27.7	0.5
Break-dancers	35.1	34.76	0.34
Exit	31.59	30.81	0.79
Race1	33.21	31.94	1.27
Rena	33.57	33.04	0.53
Average			**0.62**

Table 5-16 $PSNR_{actual}$ results using high performance method for F_3 frame

Multi-view video	With VE algorithm	Without VE algorithm	*Delta PSNR*
Akko & Kayo	29.35	29.39	-0.04
Ballroom	27.34	27.7	-0.36
Break-dancers	34.22	34.76	-0.54
Exit	30.73	30.82	-0.09
Race1	31	31.38	-0.38
Rena	32.76	32.88	-0.11
Average			**-0.25**

Table 5-17 $PSNR_{actual}$ results using conventional decimation method for F_3 frame

Multi-view video	With VE algorithm	Without VE algorithm	*Delta PSNR*
Akko & Kayo	29.55	29.47	0.08
Ballroom	27.83	27.8	0.03
Break-dancers	35	34.8	0.2
Exit	30.98	30.98	0
Race1	31.61	31.45	0.16
Rena	33.02	32.94	0.08
Average			**0.09**

Table 5-18 $PSNR_{over\text{-}estimated}$ results using conventional decimation for F_1 frame

Multi-view video	With VE algorithm	Without VE algorithm	*Delta PSNR*
Akko & Kayo	32.03	31.21	0.82
Ballroom	30.02	29.25	0.77
Break-dancers	37.49	37.03	0.46
Exit	32.71	31.71	1
Race1	34.44	32.93	1.51
Rena	37.44	36.52	0.92
Average			**0.92**

Table 5-19 $PSNR_{over\text{-}estimated}$ results using conventional decimation for F_3 frame

Multi-view video	With VE algorithm	Without VE algorithm	*Delta PSNR*
Akko & Kayo	31.74	31.6	0.13
Ballroom	29.35	29.33	0.03
Break-dancers	37.12	36.88	0.24
Exit	31.92	31.92	0
Race1	32.59	32.41	0.19
Rena	33.02	32.94	0.08
Average			**0.11**

Table 5-20 and Table 5-23 present the results using *MSSIM* video quality metric. They are consistent with the previous results, where delta quality improvement for F_1 and F_3 are 0.015 and 0.002 respectively when conventional decimation method is used for FR frames

Table 5-20 $MSSIM_{actual}$ results using high performance method for F_1 frame

Multi-view video	With VE algorithm	Without VE algorithm	*Delta PSNR*
Akko & Kayo	0.82	0.83	-0.01
Ballroom	0.8	0.8	0
Break-dancers	0.94	0.95	-0.01
Exit	0.88	0.88	0
Race1	0.91	0.91	0
Rena	0.89	0.89	0
Average			**-0.002**

Table 5-21 $MSSIM_{actual}$ results using conventional decimation method for F_1 frame

Multi-view video	With VE algorithm	Without VE algorithm	*Delta PSNR*
Akko & Kayo	0.84	0.82	0.01
Ballroom	0.83	0.8	0.03
Break-dancers	0.95	0.94	0.01
Exit	0.9	0.88	0.02
Race1	0.93	0.91	0.02
Rena	0.9	0.89	0.01
Average			**0.02**

Table 5-22 $MSSIM_{actual}$ results using high performance method for F_3 frame

Multi-view video	With VE algorithm	Without VE algorithm	*Delta PSNR*
Akko & Kayo	0.83	0.82	0.01
Ballroom	0.8	0.8	0
Break-dancers	0.94	0.94	0
Exit	0.88	0.88	0
Race1	0.9	0.9	0
Rena	0.88	0.88	0
Average			**0**

Table 5-23 $MSSIM_{actual}$ results using conventional decimation method for F_3 frame

Multi-view video	With VE algorithm	Without VE algorithm	*Delta PSNR*
Akko & Kayo	0.83	0.83	0
Ballroom	0.8	0.8	0
Break-dancers	0.94	0.94	0
Exit	0.88	0.88	0
Race1	0.91	0.91	0
Rena	0.89	0.89	0
Average			**0**

Table 5-24 and Table 5-25 summarise results using these metrics; $PSNR_{actual}$ and $MSSIM_{actual}$. From these results, applying VE algorithm with conventional decimation method provides the best visual quality among interpolation by AVC filter, regardless of the decimation method used for FR reference frame. The quality improvement is on average 0.62 dB and 0.58 dB with respect to frame interpolation by AVC filter with conventional and high performance decimation methods respectively (F_1 frame).

Visual enhancement algorithm is sensitive to the decimation method applied for FR reference frame. Conventional decimation method maintains one-to-one relationship for sub-pixel samples among FR and LR frames as shown in Figure 5-49-a. This would support direct retrieving for the corresponding samples that belong to FR reference frame (during disparity compensation). On the contrary, high performance decimation method loses this property since the sub-pixel samples are generated from samples that belong to low spatial-resolution reference frames as depicted in Figure 5-49-b.

Table 5-24 Summary results using $PSNR_{actual}$

Average $\Delta PSNR_{actual}$	**F_1**	**F_3**

PSNR VE based high performance decimation minus *PSNR* Default interpolation based high performance decimation	-0.8	-0.25
PSNR VE based conventional decimation minus *PSNR* Default interpolation based conventional decimation	0.62	0.09
PSNR VE based conventional decimation minus *PSNR* Default interpolation based high performance decimation	0.58	0.18

Table 5-25 Summary results using $MSSIM_{actual}$

Average ΔMSSIM actual	**F_1**	**F_3**
MSSIM VE based high performance decimation minus MSSIM Default interpolation based high performance decimation	-0.002	-0.001
MSSIM VE based conventional decimation minus MSSIM Default interpolation based conventional decimation	0.015	0.002
MSSIM VE based conventional decimation minus MSSIM Default interpolation based high performance decimation	0.015	0.003

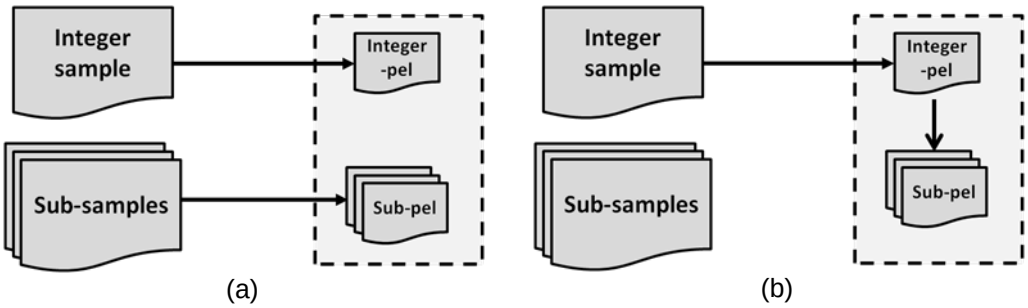


Figure 5-49 (a-b) Relation among reference frames at FR and LR using conventional and high performance decimation methods respectively

The proposed VE algorithm improves visual quality for F_1 more significantly than F_3. The amounts of inter-view prediction for both frames are analysed. The amount of inter-view predicted blocks is 81% for frames that follow Key frame (e.g. F_1), while the corresponding amount for frames that follow non key-frames (e.g. F_3) is 14% as shown in Table 5-26. Therefore, coding F_1 relies mostly on IVP, while majority of blocks that belong to F_3 are predicted by temporal frames. This explains why the proposed VE algorithm is more effective on F_1 than F_3.

Table 5-26 Amount of inter-view prediction (%) for F_1 and F_3 frames

Multi-view video	**F_1 located at T_0 [key frame]**	**F_3 located at T_1 [non-key frame]**
Akko & Kayo	91.33	13.67

Ballroom	82	10.67
Break-dancers	72.4	26.43
Exit	82.33	0.67
Race1	73.33	10.33
Rena	85.67	23.67
Average	**81.18**	**14.24**

$PSNR_{actual}$ and $MSSIM_{actual}$ metrics are extended to measure quality for coded LR frames that follow key frames. Three views are coded via H.264/AVC based multi-view video coding using first forty nine frames from each view. Sequential view prediction architecture is deployed, where each frame belongs to dependent view is predicted by nearest temporal and spatial frames. Two experiments are conducted. Coded LR frames that follow key frames are visually enhanced by the proposed VE algorithm in the 1st experiment, while these frames are interpolated by AVC 6-tap filter in the 2nd experiment. Both experiments use conventional decimation method for FR reference frames during inter-view prediction. Coded frames that follow key frames are extracted and their visual qualities are compared using these objective metrics; $PSNR_{actual}$ and $MSSIM_{actual}$. The following tables (5-27 to 5-32) provide the results for different videos. It can be seen that the proposed VE algorithm provides quality improvement than interpolation via AVC 6-tap filter through *PSNR* and *MSSIM* metrics.

Table 5-27 $PSNR_{actual}$ and $MSSIM_{actual}$ results for Akko & Kayo video

Metric	***PSNR***			***MSSIM***		
Frame	With VE algorithm	Without VE algorithm	*Delta PSNR*	With VE algorithm	Without VE algorithm	*Delta MSSIM*
F_1	29.47	29.16	0.3	0.84	0.82	0.01
F_{25}	29.35	29.03	0.32	0.83	0.82	0.01
F_{49}	29.13	28.88	0.24	0.83	0.82	0.02
F_{73}	29.61	29.24	0.37	0.84	0.82	0.02
F_{97}	29.89	29.61	0.28	0.85	0.84	0.01
F_{121}	29.64	29.26	0.38	0.85	0.84	0.02
F_{145}	29.35	29.07	0.28	0.84	0.82	0.02
Average			**0.31**			**0.02**

Table 5-28 $PSNR_{actual}$ and $MSSIM_{actual}$ results for Ballroom video

Metric	***PSNR***	***MSSIM***

Frame	With VE algorithm	Without VE algorithm	*Delta PSNR*	With VE algorithm	Without VE algorithm	*Delta MSSIM*
F_1	28.2	27.7	0.5	0.83	0.8	0.03
F_{25}	28.31	27.72	0.58	0.83	0.81	0.03
F_{49}	28.53	27.96	0.57	0.84	0.81	0.02
F_{73}	28.95	28.37	0.58	0.84	0.82	0.02
F_{97}	28.63	28.28	0.35	0.84	0.83	0.02
F_{121}	28.86	28.24	0.62	0.85	0.83	0.02
F_{145}	29.17	28.62	0.55	0.85	0.83	0.02
Average			**0.54**			**0.02**

Table 5-29 $PSNR_{actual}$ and $MSSIM_{actual}$ results for Break-dancers video

Metric	***PSNR***			***MSSIM***		
Frame	With VE algorithm	Without VE algorithm	*Delta PSNR*	With VE algorithm	Without VE algorithm	*Delta MSSIM*
F_1	35.1	34.76	0.34	0.95	0.94	0.01
F_{25}	35	34.68	0.32	0.95	0.94	0.01
F_{49}	35.03	34.8	0.24	0.95	0.95	0
F_{73}	35.15	34.99	0.16	0.95	0.95	0
F_{97}	35.1	35	0.09	0.95	0.95	0
F_{121}	35.09	34.85	0.24	0.95	0.95	0
F_{145}	35.48	35.18	0.29	0.95	0.95	0
Average			**0.24**			**0.003**

Table 5-30 $PSNR_{actual}$ and $MSSIM_{actual}$ results for Exit video

Metric	***PSNR***			***MSSIM***		
Frame	With VE algorithm	Without VE algorithm	*Delta PSNR*	With VE algorithm	Without VE algorithm	*Delta MSSIM*
F_1	31.59	30.81	0.79	0.9	0.88	0.02
F_{25}	31.9	30.98	0.92	0.9	0.88	0.02
F_{49}	31.65	30.81	0.83	0.9	0.88	0.02
F_{73}	31.79	30.91	0.88	0.9	0.88	0.02
F_{97}	31.94	30.96	0.98	0.9	0.88	0.02
F_{121}	31.89	31.16	0.74	0.9	0.88	0.02
F_{145}	31.74	31.12	0.62	0.9	0.88	0.02
Average			**0.82**			**0.02**

Table 5-31 $PSNR_{actual}$ and $MSSIM_{actual}$ results for Race1 video

Metric	***PSNR***			***MSSIM***		
Frame	With VE algorithm	Without VE algorithm	*Delta PSNR*	With VE algorithm	Without VE algorithm	*Delta MSSIM*
F_1	33.21	31.94	1.27	0.93	0.91	0.02
F_{25}	32.07	30.1	1.97	0.93	0.9	0.03
F_{49}	30.75	28.96	1.79	0.92	0.89	0.03
F_{73}	30.21	27.52	2.7	0.93	0.88	0.05
F_{97}	29.44	26.99	2.45	0.93	0.88	0.05
F_{121}	30.68	27.47	3.21	0.93	0.89	0.04
F_{145}	30.92	28	2.92	0.93	0.88	0.05
Average			**2.33**			**0.04**

Table 5-32 $PSNR_{actual}$ and $MSSIM_{actual}$ results for Rena video

Metric	***PSNR***			***MSSIM***		
Frame	With VE algorithm	Without VE algorithm	*Delta PSNR*	With VE algorithm	Without VE algorithm	*Delta MSSIM*
F_1	33.57	33.04	0.53	0.9	0.9	0
F_{25}	33.54	32.7	0.84	0.9	0.88	0.02
F_{49}	33.53	32.69	0.85	0.9	0.89	0.01
F_{73}	33.58	32.84	0.74	0.89	0.88	0.01
F_{97}	33.7	33.08	0.62	0.9	0.89	0.01
F_{121}	33.68	33.14	0.54	0.9	0.9	0
F_{145}	34.03	33.49	0.54	0.9	0.9	0
Average			**0.67**			**0.01**

The proposed VE algorithm improves visual quality of interpolated frame that follow the key frame. This improvement is due to blurriness reduction that exists in the interpolated frame. Two objective metrics have been used to measure the amount of blurriness. It involves evaluating FR frame, interpolated frames by AVC filter and the corresponding frames that are visually enhanced by the proposed VE algorithm. These metrics are *VQM* that is proposed by *Lee et al.* and $B_{average}$ component in *StSD* metric (Lee et al., 2011; De Silva et al., 2013). Multi-view video coding is used to encode three-view video, where LR frames are associated with the middle view. In each experiment, full spatial-resolution reference frames that belong to the base view are decimated by conventional method. Proposed VE algorithm is deployed on the coded LR frames that follow key frames.

The blurriness component of *StSD* metric has been used to measure the amount of blurriness in all interpolated frames that follow key frames[36]. It measures the

[36] There are seven low spatial-resolution frames that follow key frames; starting from F_1 to F_{145}

different amount of edge magnitude among un-coded frames and coded frames, where the edges are extracted by SOBEL filter. The following tables (5-33 to 5-38) show blurriness amount that is measured by $B_{average}$ (defined in *StSD* metric). It is computed for FR, LR and visually enhanced frames, where *delta* $B_{average}$ is computed by subtracting values in 4th column from values in the 3rd column. From these tables, the proposed algorithm reduces amount of blurriness for interpolated frames in the range of 0.3 to 2.9.

Table 5-33 Blurriness amount of *StSD* results for Akko & Kayo video

Time Slice	Coded FR frame	Interpolated frame without VE algorithm	Interpolated frame with VE algorithm	*Delta* $B_{average}$
T_0	6.09	8.82	7.73	1.1
T_8	6.37	9.27	8.11	1.17
T_{16}	6.07	9.11	7.66	1.45
T_{24}	6.59	9.55	8.14	1.4
T_{32}	6.32	8.94	7.46	1.49
T_{40}	6.46	9.48	7.93	1.55
T_{48}	6.05	9.53	8.02	1.51
Average				**1.38**

Table 5-34 Blurriness amount of *StSD* results for Ballroom video

Time Slice	Coded FR frame	Interpolated frame without VE algorithm	Interpolated frame with VE algorithm	*Delta* $B_{average}$
T_0	4.98	11.07	8.66	2.41
T_8	4.65	11.09	8.89	2.2
T_{16}	4.58	10.59	8.42	2.16
T_{24}	4.44	10.03	7.97	2.06
T_{32}	4.14	9.92	8.15	1.77
T_{40}	4.09	9.76	7.89	1.87
T_{48}	4.18	9.66	8.02	1.64
Average				**2.02**

Table 5-35 Blurriness amount of *StSD* results for Break-dancers video

Time Slice	Coded FR frame	Interpolated frame without VE algorithm	Interpolated frame with VE algorithm	*Delta* $B_{average}$
T_0	2.12	3.62	3.32	0.3
T_8	2.18	3.82	3.48	0.34
T_{16}	1.85	3.46	3.2	0.26
T_{24}	1.94	3.41	3.14	0.27
T_{32}	2.11	3.67	3.45	0.23
T_{40}	1.95	3.64	3.32	0.31
T_{48}	2.08	3.57	3.28	0.29

Average		**0.29**

Table 5-36 Blurriness amount of *StSD* results for Exit video

Time Slice	Coded FR frame	Interpolated frame without VE algorithm	Interpolated frame with VE algorithm	*Delta* $B_{average}$
T_0	1.69	5.69	4.83	0.86
T_8	1.6	5.5	4.51	0.99
T_{16}	1.57	5.43	4.56	0.87
T_{24}	1.66	5.43	4.51	0.92
T_{32}	1.55	5.25	4.34	0.91
T_{40}	1.55	5.07	4.24	0.83
T_{48}	1.48	5.24	4.44	0.8
Average				**0.88**

Table 5-37 Blurriness amount of *StSD* results for Race1 video

Time Slice	Coded FR frame	Interpolated frame without VE algorithm	Interpolated frame with VE algorithm	*Delta* $B_{average}$
T_0	1.51	5.3	3.46	1.83
T_8	1.13	5.85	3.63	2.22
T_{16}	1	6.36	3.77	2.59
T_{24}	0.66	6.63	3.21	3.43
T_{32}	0.39	6.72	3.46	3.26
T_{40}	0.38	6.52	3.27	3.25
T_{48}	0.42	6.6	3.15	3.45
Average				**2.86**

Table 5-38 Blurriness amount of *StSD* results for Rena video

Time Slice	Coded FR frame	Interpolated frame without VE algorithm	Interpolated frame with VE algorithm	*Delta* $B_{average}$
T_0	2.83	6.37	5.74	0.63
T_8	3.14	7	6.09	0.91
T_{16}	2.85	6.49	5.69	0.8
T_{24}	3.02	6.55	5.81	0.74
T_{32}	2.86	6.36	5.67	0.7
T_{40}	2.93	6.28	5.62	0.66
T_{48}	2.91	6	5.4	0.6
Average				**0.72**

The results are summarised in Table 5-39. The FR frames that belong to the base view have lowest blurriness, while the interpolated frames that use AVC filter have the highest blurriness. Coded FR frames suffer from blocking artefacts; therefore their edges are not significantly blurred. Interpolated frames suffer from both; blurriness and blocking artefacts. The proposed VE algorithm reduces the amount of blurriness that exists in these frames. Frames that are obtained by the proposed VE

algorithm have less blurriness than the corresponding frames that are interpolated by AVC filter. The blurriness reduction varies among different multi-view videos[37]. This is due to different scene complexities among these videos that would be represented by the Spatial Index (SI). High amount of SI indicates frames with complex details (contains many edges) while low amount of SI indicates frames with few details. E.g. for F_1 frame, SI are 33.93 and 41.61 for Break-dancers and Race1 respectively. VE algorithm would improve the visual quality for frames that contain complex details more than frames that have smooth areas.

Table 5-39 Amount of blur using blurriness component in *StSD* metric

Multi-view video	*Blur $_{average}$ (1st view)*	*Blur $_{average}$ (2nd view) without VE*	*Blur $_{average}$ (2nd view) with VE*	*Δ Blur among frames belong to 2nd view*
Akko & Kayo	6.28	9.24	7.86	1.38
Ballroom	4.44	10.3	8.29	2.02
Break-dancers	2.03	3.6	3.31	0.29
Exit	1.59	5.37	4.49	0.88
Race	0.78	6.28	3.42	2.86
Rena	2.93	6.43	5.72	0.72
Average				**1.36**

Video Quality Metric (VQM) proposed by *Lee et al*, is used to evaluate the visual quality improvement when the proposed VE algorithm is used for the interpolated frames. The metric measures the amount of *PSNR* around edges, blockiness and blurriness. Table 5-40 shows detail results using this video quality metric. FR frames have the highest *VQM* among LR frames. Visual enhanced frames by VE algorithm have fewer amounts of blockiness and blurriness artefacts than the interpolated frames by AVC filter (referred to as INT) as depicted in Table 5-41. Since visually enhanced frames inherit blocks from neighbouring FR coded frames, they have less blockiness and blurriness with respect to the corresponding coded LR frames that are interpolated by AVC filter. Table 5-42 shows amount of preserved edges. Average amount of preserved edges for frames that are visually enhanced by VE algorithm is 54.7% while the corresponding amount of frames that are interpolated by AVC filter is 28.1%. Table 5-43 summarises these results using average *VQM*, the proposed VE algorithm enhances the quality of interpolated frames, where the improvement varies from 1 to 6.9 dB.

[37] The highest and lowest improvement exist in Race1 (2.86) and Break-dancers (0.29)

Table 5-40 *VQM* $_{Lee\ et\ al.}$ comprehensive results for different videos

Multi-view video	**Akko & Kayo**	**Ballroom**	**Break-dancers**	**Exit**	**Race1**	**Rena**
VQM $_{1st\ view}$	19	20.9	26.2	27.7	30.9	26.7
VQM $_{2nd\ view\ (without\ VE)}$	8.6	6.2	13	7.4	4.8	13.1
VQM $_{2nd\ view\ (with\ VE)}$	9.7	8.8	14.5	8.6	11.7	16.3
EPSNR $_{1st\ view}$	28.1	28.4	32.3	33	34.3	33.4
EPSNR $_{2nd\ view(without\ VE)}$	22.7	21	25.6	23.1	20.4	26.8
EPSNR $_{2nd\ view(with\ VE)}$	22.2	21.2	25.4	22.1	21.9	27.7
Blocking $_{1st\ view}$	54.5	44.8	38	31.9	21.5	40.1
Blocking $_{2nd\ view(without\ VE)}$	90.2	93.2	83.4	102.1	98.2	81.7
Blocking $_{2nd\ view\ (with\ VE)}$	82.5	80.9	71.3	91.1	66.9	66.3
Blur $_{1st\ view}$	73.2	60.9	48.1	40.6	26.3	53.7
Blur $_{2nd\ view(without\ VE)}$	106.7	113.7	93.5	118.1	120.7	111.3
Blur $_{2nd\ view\ (with\ VE)}$	93	92.6	81.9	98.1	76.2	93.3

Table 5-41 Average quality improvement (*VQM* $_{Lee\ et\ al.}$) for the interpolated frames

Multi-view video	***VQM*** $_{VE}$ **–** ***VQM*** $_{INT}$	***EPSNR*** $_{VE}$ **–** ***EPSNR*** $_{INT}$	***Blocking*** $_{VE}$ **–** ***Blocking*** $_{INT}$	***Blur*** $_{VE}$ **–** ***Blur*** $_{INT}$
Akko & Kayo	1.04	-0.49	0.55	0.98
Ballroom	2.6	0.21	0.88	1.51
Break-dancers	1.51	-0.18	0.86	0.83
Exit	1.25	-0.97	0.79	1.43
Race1	6.93	1.52	2.24	3.17
Rena	3.26	0.88	1.1	1.28
Average	**2.76**	**0.16**	**1.07**	**1.53**

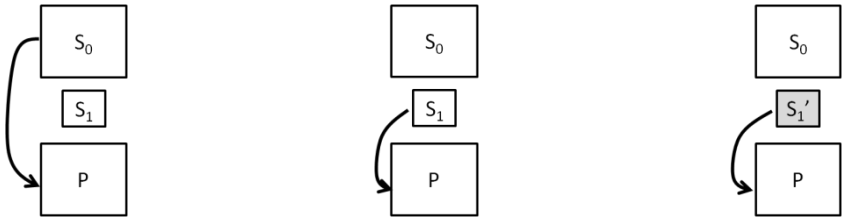
Table 5-42 Amount of preserved edges in percent via *VQM* $_{Lee\ et\ al.}$

Multi-view video	1st view (FR frames)	2nd view where coded LR frames are INT	2nd view where LR coded frames are visually enhanced
Akko & Kayo	69.1	31.7	54.3
Ballroom	77.9	30.2	53.1
Break-dancers	76.9	38.8	55.2
Exit	87	16	56.6
Race1	90.5	19.4	59.9
Rena	73.6	32.5	49.4
Average	**79.2**	**28.1**	**54.7**

Table 5-43 *VQM* average results based on *Lee et al.* proposed metric

Multi-view video	*VQM Lee et al. (1st view)*	*VQM Lee et al. (2nd view) without VE*	*VQM Lee et al. (2nd view) with VE*	*Δ VQM among frames belong to 2nd view*
Akko & Kayo	18.98	8.63	9.67	1.03
Ballroom	20.89	6.2	8.8	2.6
Break-dancers	26.2	12.99	14.5	1.51
Exit	27.75	7.38	8.63	1.25
Race	30.92	4.78	11.71	6.93
Rena	26.71	13.05	16.32	3.26
Average				**2.76**

The second application; inter-view prediction; is evaluated for mixed spatial-resolution multi-view video coding, where the middle view has LR frames. Three views are coded via H.264/AVC based multi-view video coding, where the first forty-nine frames from each view are coded. The proposed VE algorithm is used to improve the visual quality of the interpolated reference frames that belong to the second view. These frames follow key frames and they are used to predict FR frames that belong to the third view. Figure 5-50 shows three reference frame candidates to predict P-frame in the third view. These candidates are S_0, S_1 and S_1', where the last candidate is reference frame that is visually enhanced by the proposed VE algorithm. Three experiments have been conducted using different candidates for inter-view prediction. In each experiment, the statistics of inter-view predicted blocks are analysed. Table 5-44 presents the average amount of inter-view predicted blocks. From this table, the average amount of inter-view predicted blocks are increased when visually enhanced frame (S_1') is used instead of interpolated reference frame (S_1). Table 5-45 shows the coding gain when different spatial reference frames are used. When S_1' is used to predict FR frame, the bitrate is reduced by on average 33 Kbps and *PSNR* is increased by on average 0.35 dB with respect to the corresponding codec that uses S_1 reference frame. When S_0 is used instead of S_1' reference frame, the average bitrate and *PSNR* are reduced by on average 20.28 Kbps and 0.25 dB respectively. From these results, it can be implied that conducting VE algorithm for interpolated reference frames enhances the IVP when compared to the same frames that do not use the proposed VE algorithm. Full spatial-resolution frames (S_0) provide the best choice for inter-view prediction among other reference frame candidates in terms of average bitrate.



(a) (b) (c)

Figure 5-50 (a-c) Different source for inter-view prediction using FR, coded LR and visually enhanced reference frames respectively

Table 5-44 Average amount of IVP (%) using different reference frames

Multi-view video	**S_0**	**S_1**	**S_1'**
Akko & Kayo	90	86.72	88.19
Ballroom	71.9	77.43	80.05
Break-dancers	58	53.81	60.95
Exit	43.72	39.09	44.76
Race	62.62	61.62	66.43
Rena	51.51	56.29	58.13
Average	**57.55**	**62.49**	**66.42**

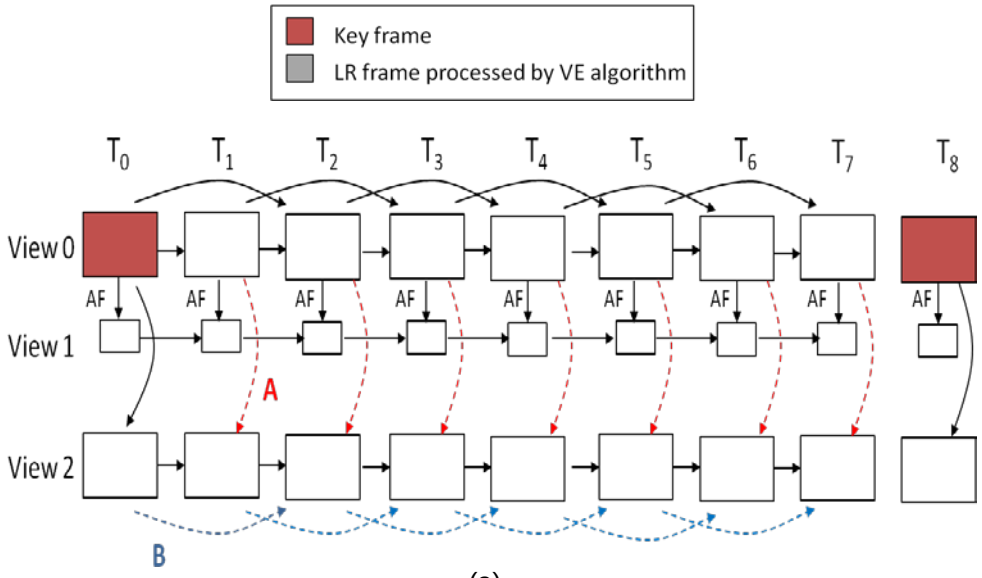
Table 5-45 Coding gain using different sources for inter-view prediction

Multi-view video	*PSNR* (dB)			Bitrate (Kbps)		
	S_0	S_1	S_1'	S_0	S_1	S_1'
Akko & Kayo	30.94	30.76	30.96	460.78	497.71	472.9
Ballroom	31.25	31.13	31.41	548.13	585.07	556.25
Break-dancers	35.3	35.2	35.55	421.80	442.19	425.51
Exit	35.55	35.56	35.8	357.31	379.59	360.87
Race1	34.71	34.62	35.51	1523.31	1716.13	1615.27
Rena	37.61	37.5	37.64	384.27	394.75	386.46

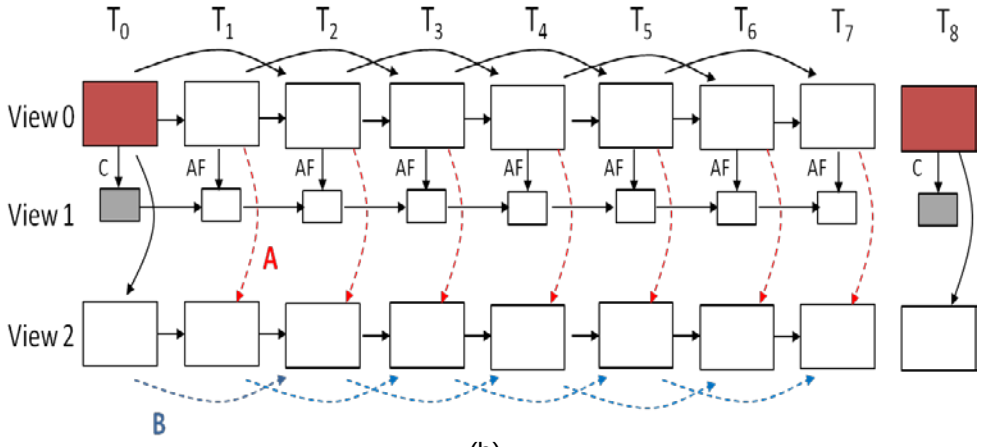
The proposed visual enhancement algorithm is deployed using two different decimation methods; conventional and high performance methods. The proposed algorithm improves visual quality for the interpolated frames, where the amount of blurriness is reduced. This is linked to the method used for decimating FR reference frames. Since conventional decimation method maintains one to one relationship among FR and LR reference frame, it provides the proposed VE algorithm the ability to use the correct samples during estimating the FR blocks that belong to the interpolated frame. The improvement is significant for frames that follow key frames, where the amount of IVP is significantly higher than the corresponding amount for the frames that follow non-key frames.

5.4.5 Proposed prediction architecture with visual enhancement algorithm

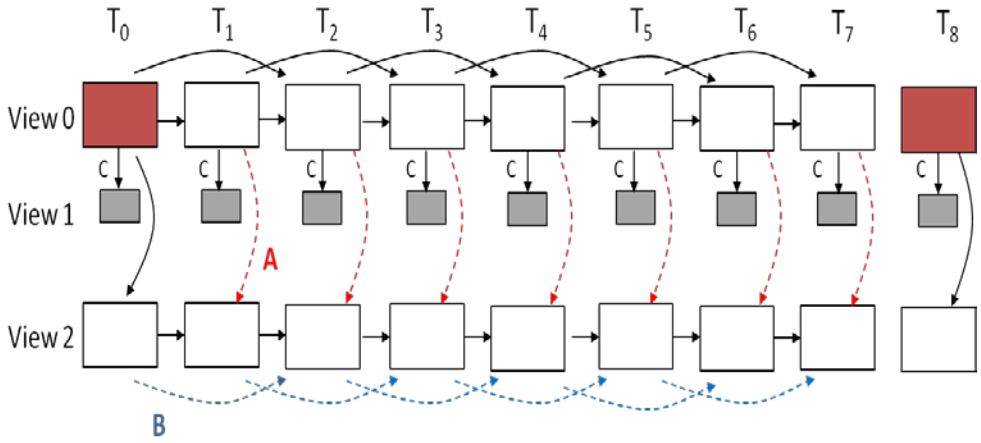
The proposed VE algorithm is integrated with the PA that is presented in subsection 5.3.4. The proposed VE algorithm is used to improve visual quality for the interpolated frames prior to display. There are three modes that correspond to different configurations within the proposed PA. Figure 5-51 shows these modes, where *A,* and *B* are two conditions that are discussed in subsection 5.3.4 (enable prediction using spatial and temporal reference frames that are referred to red and blue arrows respectively). *AF* and *C* represent high performance and conventional decimation methods respectively. Red block refers to the key frame while grey block refers to LR frame that is visually enhanced by VE algorithm.



(a)



(b)



(c)

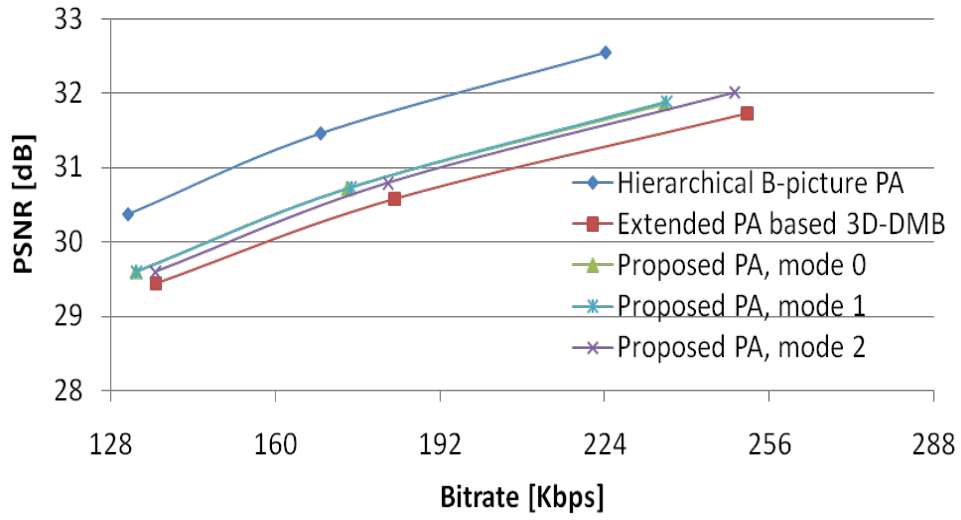
Figure 5-51 (a-c) Proposed prediction architectures using modes 0, 1 and 2

Mode 0 (Proposed MR-MVC mode 0) represents the proposed prediction architecture without integrating VE algorithm. Mode 1 (Proposed MR-MVC mode 1) integrates VE algorithm for LR frame that follows the key frame while mode 2 (Proposed MR-MVC mode 2) applies VE algorithm to all LR frames. These modes provide trade-off among visual quality and bitrate. Mode 1 changes only the decimation method for FR reference frames (key frames). Mode 2 applies two changes in the proposed prediction architecture, where temporal prediction for low spatial-resolution frames is omitted. This would improve the visual quality for all interpolated frames that would rely on the inter-view prediction. Also all FR frames that belong to the base view are decimated by the conventional method.

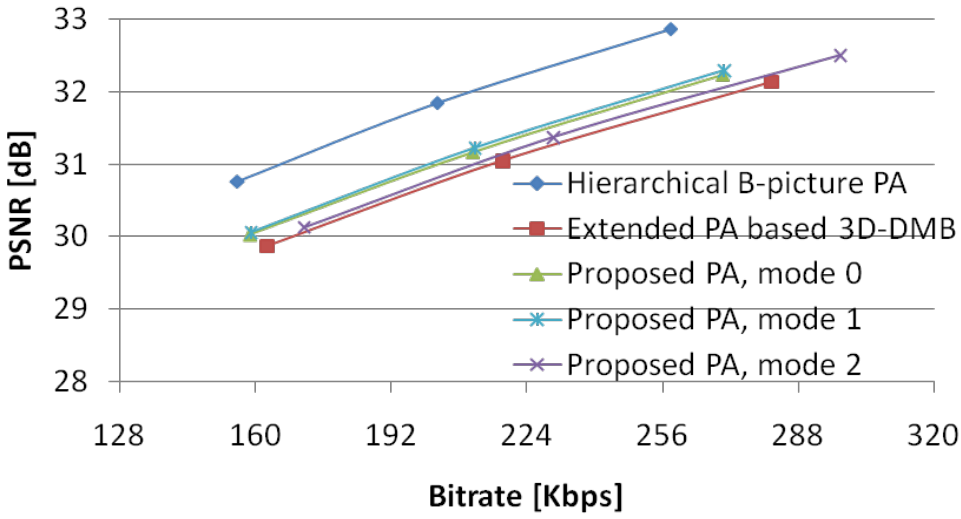
5.4.6 Results and discussions

The proposed prediction architecture among three modes is evaluated alongside the extended architecture based 3D-DMB and HBP. Figure 5-52 shows rate-distortion curves when different prediction architectures are used, where Y-axis and X-axis are $PSNR_{actual}$ and bitrate respectively.

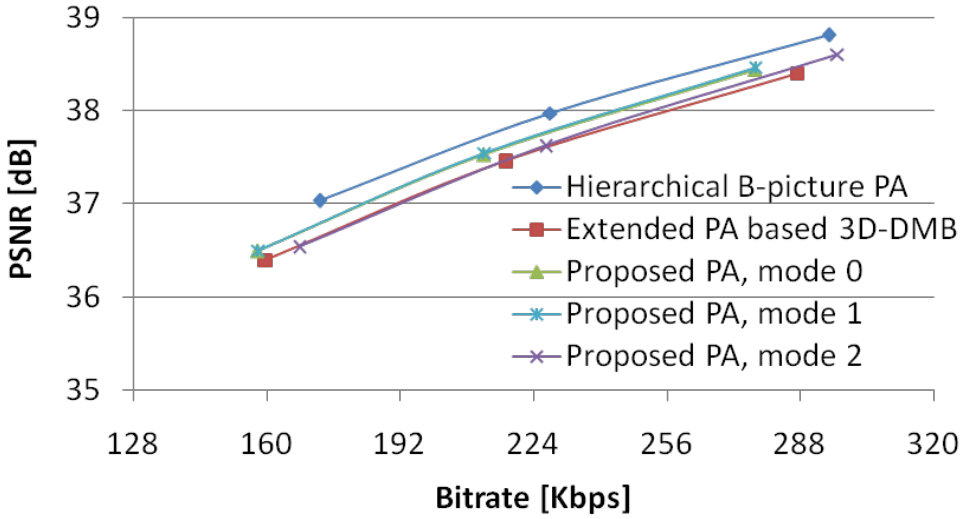
Table 5-46 and Table 5-47 provide the results when coding mixed spatial-resolution multi-view videos using the proposed prediction architecture, HBP and the extended architecture based 3D-DMB. From these results, VE algorithm improves visual quality for the interpolated frames at the expense of increasing average bitrate (e.g. the proposed PA using mode 2 processes VE algorithm for all LR frames by omitting temporal prediction for the view that contains LR frames).


33
32
31
30
29
28
PSNR [dB]
128
160
192
224
256
288
Bitrate [Kbps]
Hierarchical B-picture PA
Extended PA based 3D-DMB
Proposed PA, mode 0
Proposed PA, mode 1
Proposed PA, mode 2

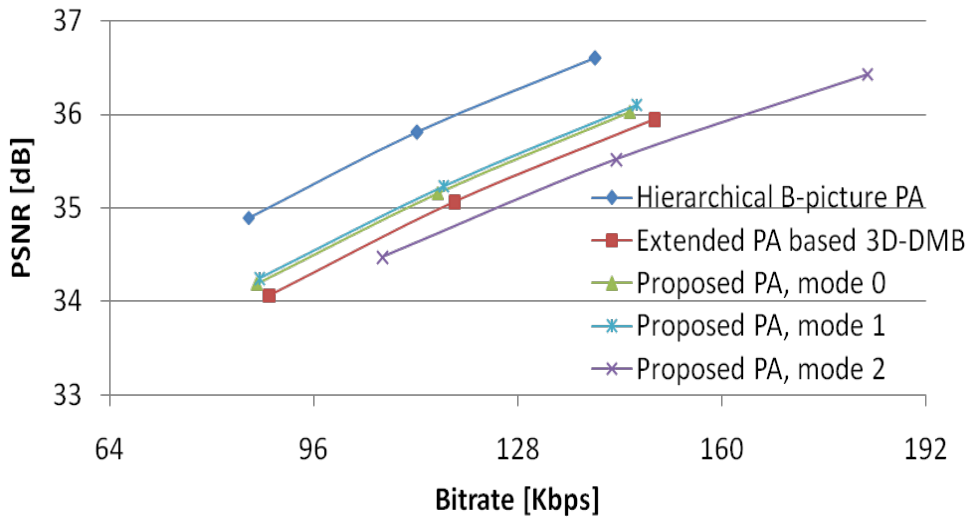
(a)


33
32
31
30
29
28
PSNR [dB]
128
160
192
224
256
288
320
Bitrate [Kbps]
Hierarchical B-picture PA
Extended PA based 3D-DMB
Proposed PA, mode 0
Proposed PA, mode 1
Proposed PA, mode 2

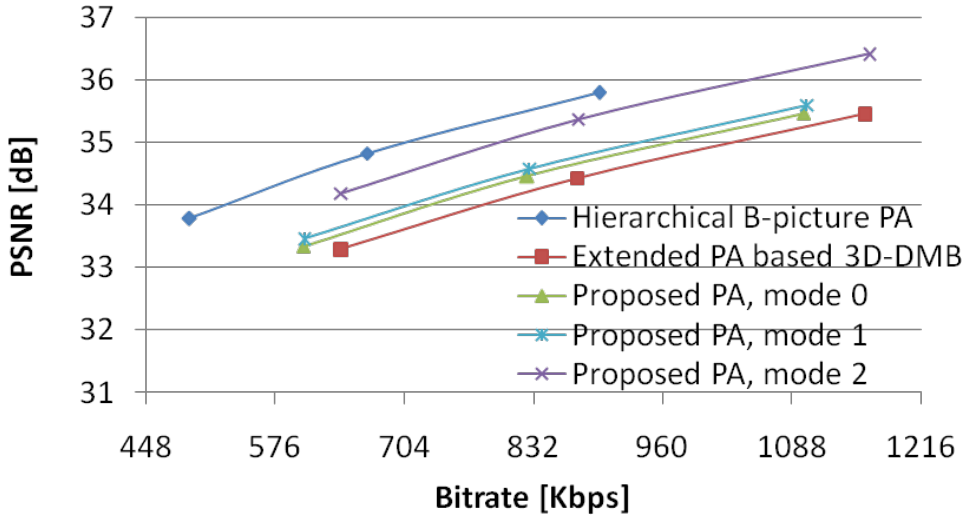
(b)


39
38
37
36
35
PSNR [dB]
128
160
192
224
256
288
320
Bitrate [Kbps]
Hierarchical B-picture PA
Extended PA based 3D-DMB
Proposed PA, mode 0
Proposed PA, mode 1
Proposed PA, mode 2

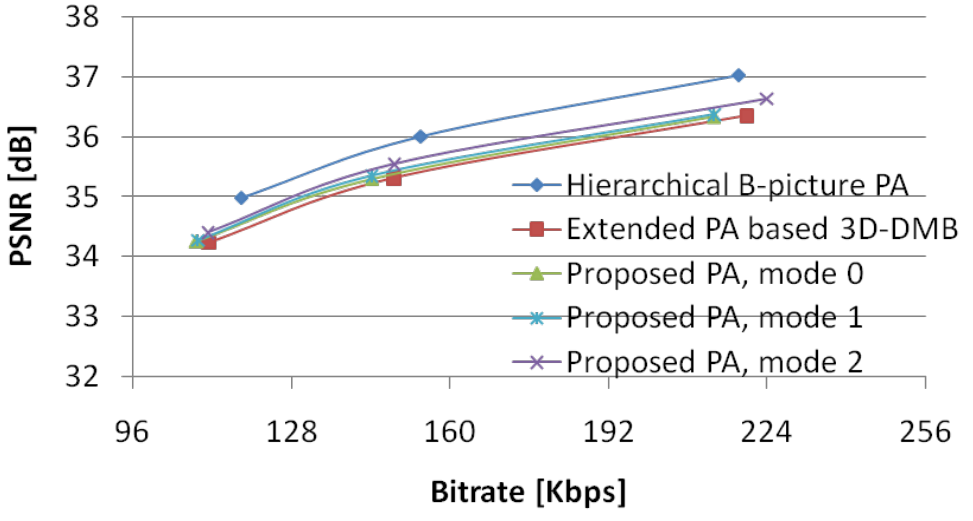
(c)



(d)



(e)



(f)

Figure 5-52 Rate-distortion curves for the proposed prediction architecture among different modes

Table 5-46 ∆*PSNR* results for proposed prediction architecture

ΔPSNR (dB)	*PSNR(Mode i) – PSNR(extended PA based 3D-DMB)*			*PSNR(Mode i) – PSNR(HBP)*		
Mode	**0**	**1**	**2**	**0**	**1**	**2**
Akko & Kayo	0.13	0.15	0.21	-0.75	-0.73	-0.67
Ballroom	0.13	0.18	0.32	-0.68	-0.63	-0.48
Break-dancers	0.07	0.08	0.17	-0.45	-0.43	-0.35
Exit	0.09	0.17	0.45	-0.65	-0.57	-0.29
Race1	0.02	0.15	0.94	-0.37	-0.25	0.54
Rena	-0.01	0.04	0.24	-0.71	-0.66	-0.47
Average	**0.07**	**0.13**	**0.39**	**-0.6**	**-0.54**	**-0.29**

Table 5-47 ∆bitrate results for proposed prediction architecture

ΔBR (Kbps)	BR(Mode i) – BR(extended PA based 3D-DMB)			BR(Mode i) – BR(HBP)		
Mode	**0**	**1**	**2**	**0**	**1**	**2**
Akko & Kayo	-9.07	-8.34	-1.2	5.26	5.99	13.13
Ballroom	-7.2	-6.5	11.89	8.18	8.88	27.27
Break-dancers	-5.5	-5.34	9.68	-16.08	-15.92	-0.9
Exit	-2.68	-1.8	25.32	3.23	4.11	31.23
Race1	-49.88	-47.31	1.32	158.67	161.24	209.87
Rena	-4.4	-4.42	0.2	-9.89	-9.91	-5.29
Average	**-13.12**	**-12.29**	**7.87**	**24.9**	**25.73**	**45.89**

The summary results in terms of *ΔPSNR* and ΔBR are presented in Table 5-48 and Table 5-49 respectively. From these results, deploying mode 1 has slightly improved actual *PSNR* while average bitrate increases by 0.84 Kbps with respect to mode 0. Mode 2 provides the highest visual quality for the proposed PA with respect to other modes. Its improvements are on average 0.39 dB and -0.29 dB at the expense of increasing average bitrate by 7.87 Kbps and 45.89 Kbps with respect to deploying extended architecture based 3D-DMB and HBP prediction architectures respectively. The proposed VE algorithm increases coding time by 0.038% and 0.049% for modes 1 and 2 respectively. Therefore the computational complexity for the proposed VE algorithm is not considered a burden for MVC.

Table 5-48 *ΔPSNR* summary results

ΔPSNR $_{actual}$	*PSNR(Proposed Prediction architecture) – PSNR (Extended PA based 3D-DMB / HBP architectures)*		
Prediction architecture	**Mode 0**	**Mode 1**	**Mode 2**
Extended PA based 3D-DMB	0.072	0.128	0.387
HBP	-0.601	-0.545	-0.285

Table 5-49 ΔBitrate summary results

ΔBR	BR(Proposed Prediction architecture) – BR (Extended 3D-DMB / HBP architecture)		
Prediction architecture	**Mode 0**	**Mode 1**	**Mode 2**
Extended architecture based 3D-DMB	-13.122	-12.285	7.868
HBP	24.895	25.732	45.885

Tables 5-50 to 5-52 show $PSNR_{actual}$ results when coding the middle view that contains LR frames by different prediction architectures. Deploying VE algorithm for all interpolated frames improves visual quality of these frames. The quality improvements are on average 0.9 dB and 0.6 dB with respect to the corresponding interpolated frames that are coded by the extended architecture based 3D-DMB and HBP prediction architectures respectively. VE algorithm is deployed in two modes, where mode 1 executes the proposed VE algorithm for a set of LR frames that follow key frames. The average quality improvement for the view that has LR frames using mode 1 is 0.2 dB at the expense of slightly increasing the average bitrate by 0.84 Kbps with respect to mode 0. Mode 2 allows further visual quality improvement, where the average quality gain for the view that has LR frames is 0.9 dB and the average bitrate is increased by 21 Kbps with respect to mode 0.

Table 5-50 $PSNR_{actual}$ results for interpolated frames

Prediction architecture	Akko & Kayo	Ballroom	Break-dancers	Exit	Race1	Rena
Extended 3D-DMB	29.13	28.189	35.021	30.944	28.263	32.778
HBP	29.615	28.589	35.288	31.164	28.238	33.271
mode 0	29.133	28.18	35.006	30.942	28.249	32.769
mode 1	29.191	28.315	35.049	31.165	28.609	32.924
mode 2	29.356	28.732	35.3	31.914	30.893	33.497

Table 5-51 *ΔPSNR* for interpolated frames with respect to extended PA based 3D-DMB

ΔPSNR (dB): PSNR(proposed architecture) – PSNR(extended PA based 3D-DMB)			
Proposed prediction architecture with	Mode 0	Mode 1	Mode 2
Akko & Kayo	0.003	0.061	0.226
Ballroom	-0.01	0.125	0.543
Break-dancers	-0.015	0.028	0.279
Exit	-0.002	0.221	0.97
Race1	-0.014	0.346	2.63
Rena	-0.009	0.146	0.718
Average	**-0.008**	**0.154**	**0.894**

Table 5-52 *ΔPSNR* for interpolated frames with respect to HBP architecture

ΔPSNR (dB): PSNR(proposed architecture) – PSNR(HBP)			
Proposed prediction architecture with	Mode 0	Mode 1	Mode 2
Akko & Kayo	-0.482	-0.424	-0.259
Ballroom	-0.409	-0.275	0.143
Break-dancers	-0.282	-0.239	0.012
Exit	-0.222	0.001	0.751
Race1	0.011	0.37	2.655
Rena	-0.502	-0.347	0.225
Average	**-0.314**	**-0.152**	**0.588**

5.4.7 Conclusions

A visual enhancement algorithm has been proposed that improves visual quality for coded LR frames. During disparity compensation, the blocks that belong to FR frames are used among the interpolated residual to substitute blocks that belong to the interpolated frames. The VE algorithm would be used in display and enhancing inter-view prediction. The former application targets reducing blurriness, while the latter improves visual quality for the interpolated reference frames prior to conducting disparity estimation. A set of modes have been presented to provide different trade-off among visual quality for the interpolated frames and average bitrate. Processing the proposed VE algorithm for the interpolated frames through the proposed PA, mode 2 provides the highest visual quality improvement among corresponding frames that are coded by HBP and the extended PA based 3D-DMB. The quality improvements for these frames are on average 0.9 dB and 0.6 dB at the expense of increasing bitrate by on average 8 Kbps and 46 Kbps with respect to extended PA based 3D-DMB and HBP PA respectively.

5.5 Summary of the investigations

This chapter investigated mixed spatial-resolution multi-view video coding at low bitrates. First, it discussed how much inter-view prediction is affected when coding frames with different spatial-resolution. Deploying a FR frame as reference frame provides better coding efficiency than using a LR frame when coding mixed spatial-resolution stereoscopic video, by on average 0.63 dB while saving bitrate by 6.2%. When asymmetric quality is deployed with mixed spatial-resolution stereoscopic video coding that deploys LR frames in a base view, 44% of the variation in the IVP can be explained by asymmetric quality, according to regression analysis. The

relationship of IVP and ΔQP using six multi-view videos could be described by equation 5-1.

Different methods for decimation and interpolation of reference frames are compared. High performance methods are recommended for decimation and interpolation since they achieve similar coding gain and less time for filtering compared to conventional methods. This is due to deploying filtering on fewer samples than conventional methods.

Statistical analysis of block matching is then applied for low and full spatial-resolution frames. Recent temporal and spatial FR reference frames have most significant contribution of block matching when coding FR and LR frames. Through analysing the correlation among temporal and inter-view predicted blocks during coding neighbouring frames, spatial and 2nd temporal reference frames are used when their expected role of block matching are significant. This is beneficial when coding multi-view video that contains large disparities and slow objects motion. Based on the previous results, prediction architecture is proposed and evaluated among HBP and extended architecture based 3D-DMB. The proposed prediction architecture saves a significant amount of memory required for DPB by 51.9% and 31.6% with respect to HBP and extended architecture based 3D-DMB respectively. The proposed prediction architecture accelerates encoding by on average 57% and up to 77.5% with respect to the corresponding time needed by hierarchical B-picture architecture. It speeds up encoding by on average 14% and up to 54% with respect to an extended prediction architecture based 3D-DMB. The proposed PA needs less bitrate for coding asymmetric MVV by on average 13.1 Kbps with respect to extended architecture based 3D-DMB, while both obtain similar quality for decoded MVV. HBP PA provides higher coding efficiency than the proposed PA, where HBP PA obtains better quality by on average 0.78 dB while requiring less bitrate by on average 24.9 Kbps with respect to the proposed PA.

The proposed PA with adaptive reference frame ordering algorithm saves on average 28.7 Kbps and 35.4 Kbps with respect to an HBP architecture and extended architecture based on 3D-DMB, respectively. It provides a similar quality for decoded asymmetric MVV to the corresponding video coded via extended architecture based 3D-DMB. HBP provides better quality by on average 0.38 dB compared to the corresponding video that is coded by the proposed prediction architecture. The proposed prediction architecture accelerates compression time by on average 64% and 33% with respect to the corresponding time needed by HBP architecture and the extended PA based 3D-DMB.

A visual enhancement algorithm has been proposed to reduce the amount of blurriness that exists in coded LR frames. During disparity compensation, the blocks that belong to FR frames are used among interpolated residuals to substitute blocks that belong to the interpolated frames. Different modes have been presented to provide a trade-off among visual quality for the interpolated frames and average bitrate. Integrating VE algorithm for the interpolated frames (mode 2) provides the highest visual quality improvement among corresponding frames that are coded by other prediction architectures. The quality improvements for the interpolated frames are on average 0.9 dB and 0.6 dB at the expense of increasing bitrate by on average 8 Kbps and 46 Kbps with respect to extended architecture based 3D-DMB and HBP prediction architectures respectively.

The next chapter will summarise the outcomes of the research investigations that are reported in the thesis, followed by the research directions that could be addressed in the future.

CHAPTER 6. CONCLUSIONS AND FUTURE WORK

This chapter presents the outcomes of the research investigations and outlines several research directions that could be addressed in the future.

6.1 Conclusions of research investigations

The following outcomes of the research investigations for symmetric multi-view video coding are summarised as follows:

- The camera separation affects the coding performance for multi-view video coding. Although increasing the camera separation reduces coding efficiency for MVC, it cannot be used as a reliable criterion when selecting a suitable coding solution for a given multi-view video. Scene complexity affects inter-camera angle threshold, where datasets with a dominant temporal correlation have a lower threshold than datasets with balanced correlations among spatial and temporal frames.
- Prediction architectures have been investigated in terms of RFS and RFO. Based on the block matching analysis, the nearest two frames in temporal, spatial and spatiotemporal directions are chosen for RFS. Interleaved RFO is more consistent with the block matching analysis than other static reference frame ordering. The proposed prediction architecture achieves a superior coding performance relative to other architectures by a coding gain up to 2.3 dB. Since few reference frames have the majority of block matching contributions using a subset of coding modes, a trade-off study among coding efficiency and computational complexity was conducted. For low complexity multi-view video codec, the nearest temporal and spatial frames are used for reference frame selection, while macroblock partitions coding modes are enabled.
- Adaptive reference frame ordering algorithm is proposed, where RFO for the current frame is predicted by analysing block matching statistics for recent temporal frame. When the scene changes, reference frames indices are reordered in a way that places the spatial reference frame first rather than the temporal reference frame in List 0. The algorithm has been tested in two applications: through coding multi-view videos using multiple reference frames, and compressing a sequence that contains hard scene changes. For prediction architectures with multiple reference frames, the algorithm improves the coding gain for the codec by up to 0.2 dB. When coding a sequence that contains multiple

scenes, the algorithm saves bitrate by up to 6.2% with respect to a prediction architecture that deploys a static reference frame ordering.

In context of mixed spatial-resolution multi-view video coding investigations, the following outcomes are summarised as follows:

- The first study explores the effect of inter-view prediction direction on the coding performance of mixed spatial-resolution stereoscopic video coding. Deploying FR rather than LR frames in the base view achieves a higher coding gain by on average 0.63 dB while the bitrate is reduced by 6.2%. The results published by *Brust et al.* regarding the effect of different inter-view prediction directions on coding performance of stereoscopic video coding are biased to asymmetric quality (Brust et al., 2010). Based on regression analysis for asymmetric quality, and mixed spatial-resolution stereoscopic video coding using six multi-view videos, the relationship of Inter-View Prediction (IVP) and ΔQP would be described by the equation: $IVP = 1.492 + 1.096\,\Delta QP$
- Different decimation and interpolation methods have been evaluated in terms of coding gain and computational complexity. High performance methods for decimation and interpolation have similar coding gain and require less computational complexity than the conventional methods. This is due to the deployment of filtering to less number of samples than the conventional methods. Conventional decimation and interpolation methods maintain a one-to-one relationship among samples at full and low spatial-resolution, in contrast to high performance methods.
- The prediction architecture has been defined by statistical analysis of block matching among candidate reference frames. Nearest temporal and spatial FR reference frames are used during coding of full and low spatial-resolution frames. Spatial and second temporal reference frames are selected when their expected amount of block matching are significant during coding FR frame that belong to the dependent view. Based on block matching statistics results, prediction architecture is proposed and evaluated among HBP and extended architecture based 3D-DMB. The proposed PA reduces DPB size by 51.9% and 31.6% with respect to HBP and extended architecture based 3D-DMB respectively. The proposed prediction architecture speeds-up encoding by on average 57% and 14% with respect to the corresponding time needed by HBP and extended architecture based 3D-DMB respectively. The proposed prediction architecture needs less bitrate for coding asymmetric multi-view video by on average 13.1

Kbps with respect to the extended architecture based 3D-DMB. HBP architecture is more coding efficient than the proposed architecture, where it obtains better quality by on average 0.78 dB while requiring less bitrate by on average 24.9 Kbps.

- Adaptive reference frame ordering algorithm has been integrated with the proposed PA. It saves bitrate by on average 28.7 Kbps and 35.4 Kbps with respect to HBP architecture and extended architecture based 3D-DMB, respectively. The proposed prediction architecture speeds up encoding by on average 64% and 33% with respect to the corresponding time needed by hierarchical B-picture architecture and the extended PA based 3D-DMB.
- A visual enhancement algorithm is proposed to improve visual quality for the interpolated frames that utilise the information derived from disparity compensation. Blocks that belong to the interpolated frame are substituted by summation of predicted blocks that belong to the FR reference frame and the interpolated signals from residuals. The algorithm is sensitive to decimation method that is deployed to FR reference frame during inter-view prediction. Frames processed by the algorithm have higher visual quality than the corresponding frames that are interpolated by an AVC filter. This is conditional to deploying the conventional decimation method for FR frames. The improvement is more significant for the interpolated frames that follow key frames rather than frames that follow non-key frames. This is due to significant amount of inter-view prediction of former frames. The visual quality improvement is validated using *PSNR*, *MSSIM*, Blurriness component of *StSD* and *VQM* proposed by *Lee et al*. metric. Different modes have been presented to provide trade-off among visual quality of the interpolated frames and average bitrate. Integrating the VE algorithm for all frames (mode 2) gets the highest visual quality improvement among corresponding frames that are coded by other prediction architectures. The quality improvements for interpolated frames are on average 0.9 dB and 0.6 dB at the expense of increasing bitrate by on average 8 Kbps and 46 Kbps with respect to the extended architecture based 3D-DMB and HBP prediction architectures respectively.

In summary, the research investigated the impact of camera separation and prediction architectures in context of symmetric MVC. Inter-camera angle as standalone criteria is not sufficient to decide the best use for MVC. Through conducting statistical analysis of block matching, prediction architectures are proposed in addition to proposing adaptive reference frame ordering algorithm that

is beneficial when coding videos with hard scene changes. In context of mixed spatial-resolution MVC, several studies are tackled towards deriving prediction architecture. First, the impact of inter-view prediction direction is studied then different decimation and interpolation methods are examined in addition to conducting block matching statistics. The proposed prediction architecture provides comparable coding performance, consumes less computational complexity and memory size than other prediction architectures that are common used through this coding approach. Visual enhancement is tackled for the interpolated frames. Low computational complexity solution is proposed, where the information embedded in disparity compensation is used to reduce the amount of blurriness in the interpolated frame at the receiver side.

Parts of the outcomes that have been reported in the thesis are published that include the investigations reported in sections 4.1, 4.3 and 4.4. The papers are attached in the publications section.

6.2 Future work

There are several research directions that could be addressed in future work. The following summarise these research directions:

- The proposed visual enhancement algorithm needs further improvement. Since the proposed algorithm does not apply for all blocks, boundaries among intra/temporal blocks and inter-view predicted blocks might be visible. One of the candidate solutions is applying a Deblocking filter, where the pixels related to both blocks are filtered by different weights.
- The proposed visual enhancement algorithm provides a low-complexity solution for interpolated frames at the expense of a bitrate increase. The super-resolution by example-based method could improve visual quality for the interpolated frames without increasing the bitrate at the expense of high computational complexity. This needs to be investigated to compare both methods in terms of rate-distortion and computational complexity.
- *Jain et al.* proposed alternate blur format for mixed spatial-resolution stereoscopic video coding (Jain et al., 2014). Their proposed format reduces the amount of eye fatigue relative to a single-blur format especially for animated scenes. The proposed visual enhancement algorithm could be applied to an alternate blur format, where an interpolated frame will use information from temporal and disparity compensations to improve its visual quality. Objective and subjective

assessments are necessary to compare both single-blur and alternate blur formats for mixed spatial-resolution multi-view video coding.

- Mixed spatial-resolution multi-view video coding could be deployed in the context of multi-view plus depth. Texture and depth maps among neighbouring views could have different spatial-resolution. This could further reduce the bitrate compared to deploying each coding approach separately. At the decoder side, the interpolated frames could be visually improved by the proposed visual enhancement algorithm while the frames belonging to intermediate views are synthesised.

BIBLIOGRAPHY

Abdoli, M., Soryani, M. & Modarres, a. F.A., The Impact of View Spacing in Multi-view Video Compression Efficiency. In: *Seventh International Conference on Information Technology: New Generations*. 2010, pp. 1314–1315.

Aflaki, P., Hannuksela, M.M. & Gabbouj, M. (2013). Subjective Quality Assessment of Asymmetric Stereoscopic 3D Video. *Signal, Image and Video Processing*, pp. 1–15.

Aflaki, P., Hannuksela, M.M. & Gabbouj, M. (2014). Adaptive Spatial Resolution Selection for Stereoscopic Video Compression with MV-HEVC: A Frequency Based Approach. In: *2014 IEEE International Symposium on Multimedia*. pp. 267–270.

Aflaki, P., Hannuksela, M.M., Hakkinen, J., Lindroos, P. & Gabbouj, M., Subjective study on compressed asymmetric stereoscopic video. In: *IEEE International Conference on Image Processing*. 2010, pp. 4021–4024.

Aflaki, P., Hannuksela, M.M., Homayouni, M. & Gabbouj, M., Cross-asymmetric Mixed-resolution 3D Video Compression. *3DTV-Conference*. 2012, pp. 1–4.

Aflaki, P., Rusanovskyy, D., Utriainen, T., Pesonen, E. & Hannuksela, M.M., Study of Asymmetric Quality between Coded Views in Depth- Enhanced Multiview Video Coding. In: *International Conference on 3D Imaging*. 2011, p. 8.

Aflaki, P., Su, W., Joachimiak, M., Rusanovskyy, D. & Hannuksela, M.M., Coding of Mixed-Resolution Multiview Video in 3D Video Application. In: *ICIP*. 2013, pp. 1704–1708.

Ahn, Y., Hwang, T., Sim, D. & Han, W. (2014). Implementation of Fast HEVC Encoder Based on SIMD and Data-level Parallelism. *EURASIP Journal on Image and Video Processing*. (1). pp. 1–19.

Aksay, A., Bilen, C., Kurutepe, E., Ozcelebi, T., Akar, G.B., Civanlar, M.R. & Tekalp, A.M., Temporal and Spatial Scaling for Stereoscopic Video Compression. In: *EUSIPCO*. 2006, pp. 1–5.

Alatan, A.A., Yemez, Y., Gudukbay, U., Zabulis, X., Muller, K., Erdem, C.E., Weigel, C. & Smolic, A. (2007). Scene Representation Technologies for 3DTV—A Survey. *IEEE Transactions on Circuits and Systems for Video Technology*. 17 (11). pp. 1587–1605.

An, P., Guo, Q., Mi, T., Zhou, L. & Zhang, Z., Multi-view Video Coding Based on View Prediction. In: *International Conference on Audio, Language and Image Processing*. 2008, pp. 1481–1485.

Antipolis, S. (2005). *TS 102 428 - V1.1.1* - Digital Audio Broadcasting (DAB); DMB Video Service; User Application Specification, report, Available from: http://www.etsi.org/deliver/etsi_ts/102400_102499/102428/01.02.01_60/ts_102428v010201p.pdf

Bal, C. (2009). Three-dimensional Video Coding on Mobile Platforms*, thesis,* Available from: www.thesis.bilkent.edu.tr/0003940.pdf

Barsi, A., Balogh, T., Nagy, Z. & Kovacs, P. D5.1-Requirements and specifications for 3D video, *3DPHONE Project no. FP7-213349*. 2008, p.20.

Belloulata, K. & Zhu, S. (2007). A New Object-Based System for Fractal Video Sequences Compression. *Journal of Multimedia*. 2 (3). pp. 17–25.

Benzie, P., Watson, J., Surman, P., Rakkolainen, I., Hopf, K., Urey, H., Sainov, V. & von Kopylow, C. (2007). A Survey of 3DTV Displays: Techniques and Technologies. *IEEE Transactions on Circuits and Systems for Video Technology*. 17 (11). pp. 1647–1658.

Bilen, C., Aksay, A. & Akar, G., A Multi-View Video Codec Based on H.264. In: *ICIP Conference*. 2006, pp. 541–544.

Boev, A., Gotchev, A., Poikela, M. & Aksay, A. (2011). Modelling of the Stereoscopic HVS, *MOBILE3DTV Project no. 216503*. 2011, p. 68.

Boev, A., Hollosi, D. & Gotchev, A., Classification of Stereoscopic Artefacts*, MOBILE3DTV Project no. 216503*. 2011, p. 59.

Bosc, E., Jantet, V., Pressigout, M., Morin, L. & Guillemot, C., Bit-Rate Allocation for Multi-view Video Plus Depth. In: *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*. 2011, pp. 3–6.

Bossen, F., Bross, B., Suhring, K. & Flynn, D. (2012). HEVC Complexity and Implementation Analysis. *IEEE Transactions on Circuits and Systems for Video Technology*. 22 (12). pp. 1685–1696.

Bourge, A., Gobert, J. & Bruls, F., MPEG-C Part 3: Enabling the Introduction of Video Plus Depth Contents. In: *IEEE Workshop on Content Generation and Coding for 3D-television*. 2006, pp. 3–6.

Bouyagoub, S., Sheikh Akbari, A., Bull, D. & Canagarajah, N. (2010). Impact of Camera Separation on Performance of H.264/AVC-based Stereoscopic Video Codec. *Electronics Letters*. 46 (5). p. 345.

Brandt, J., Trotzky, J. & Wolf, L., Fast Frame-Based Scene Change Detection in the Compressed Domain for MPEG-4 Video. In: *The Second International Conference on Next Generation Mobile Applications, Services, and Technologies*. 2008, pp. 514–520.

Brust, H., Smolic, A., Mueller, K., Tech, G. & Wiegand, T., Mixed resolution Coding of Stereoscopic Video for Mobile Devices. In: *3DTV Conference*. 2009, pp. 1–4.

Brust, H., Tech, G., Mueller, K. & Wiegand, T., Mixed Resolution Coding with Inter view Prediction for Mobile 3DTV. In: *3DTV Conference.* 2010, pp. 1–4.

Chen, X. & Luthra, A., MPEG-2 Multiview Profile and its Application in 3D TV. In: S. Panchanathan & F. Sijstermans (eds.). *SPIE Conference*. 1997, pp. 212–223.

Chen, Y. & Vetro, A. (2014). Next-generation 3D Formats with Depth Map Support. *IEEE Multimedia*. 21 (2). pp. 90–94.

Chen, Y., Liu, S., Wang, Y., Hannuksela, M.M., Li, H. & Gabbouj, M., Low-Complexity Asymmetric Multiview Video Coding. In: *ICME Conference.* 2008, pp. 773–776.

Chen, Y., Wang, Y., Hannuksela, M.M. & Gabbouj, M., Picture-level Adaptive Filter for Asymmetric Stereoscopic Video. In: *ICIP Conference*. 2008, pp. 1944–1947.

Chen, Y., Wang, Y.-K., Gabbouj, M. & Hannuksela, M.M., Regionally Adaptive Filtering for Asymmetric Stereoscopic Video Coding. In: *IEEE International Symposium on Circuits and Systems*. 2009, pp. 2585–2588.

Chen, Y., Wang, Y.-K., Ugur, K., Hannuksela, M.M., Lainema, J. & Gabbouj, M. (2009). The Emerging MVC Standard for 3D Video Services. *EURASIP Journal on Advances in Signal Processing*. (1). p. 13.

Chiang, J., Chen, W., Liu, L., Hsu, K. & Lie, W. (2011). A Fast H.264 / AVC-Based Stereo Video Encoding Algorithm Based on Hierarchical Two-Stage Neural Classification. *IEEE Journal of selected topics in signal processing*. 5 (2). pp. 309–320.

Chiang, J.C., Hou, P.H., Liu, K.C. & Lie, W.N., Multiview Texture Coding and Free Viewpoint Image Synthesis for Mesh-based 3D Video Transmission. In: *ISCAS Conference*. 2012, pp. 377–380.

Choi, M., Kim, J., Cho, W. & Burm, J., Area-Efficient Fast Scheduling Schemes for MVC Prediction Architecture. In: *IEEE International Symposium on Circuits and Systems*. 2011, pp. 575–578.

Chung, T., Song, K. & Kim, C.-S., Efficient Prediction Structure for 2-D Wide Multi-view Video Sequence. In: *International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. 2008, pp. 1243–1246.

Chung, T.-Y., Jung, I.-L., Song, K. & Kim, C.-S. (2010). Multi-view Video Coding with View Interpolation Prediction for 2D Camera Arrays. *Journal of Visual Communication and Image Representation*. 21 (5-6). pp. 474–486.

Daribo, I. & Saito, H. (2011). A Novel Inpainting-based Layered Depth Video for 3DTV. *IEEE Transactions on Broadcasting*. 57 (2). pp. 533–541.

De Silva, V., Arachchi, H.K., Ekmekcioglu, E. & Kondoz, A. (2013). Toward an Impairment Metric for Stereoscopic Video: a Full-reference Video Quality Metric to Assess Compressed Stereoscopic Video. *IEEE Transactions on Image Processing*. 22 (9). pp. 3392–404.

De Silva, V., Arachchi, H.K., Ekmekcioglu, E., Fernando, A., Dogan, S., Kondoz, A. & Savas, S., Psycho-physical Limits of Interocular Blur Suppression and its Application to Asymmetric Stereoscopic Video Delivery. In: *19th International Packet Video Workshop*. 2012, pp. 184–189.

Dodgson, N.A. (2005). Autostereoscopic 3D displays. *Computer*. 38 (8). pp. 31–36.

Dufaux, F., Pesquet-Popescu, B. & Cagnazzo, M. (2013). Emerging Technologies for 3D Video. *Chichester, UK: John Wiley & Sons, Ltd Publication.*

Eichhorn, a. & Ni, P., Pick Your Layers Wisely - A Quality Assessment of H.264 Scalable Video Coding for Mobile Devices. In: *IEEE International Conference on Communications*. 2009, pp. 1–6.

Eisert, P. (2000). Very Low Bit-Rate Video Coding Using 3-D Models, thesis, Available from: https://www.informatik.hu-berlin.de/de/forschung/gebiete/viscom/papers/diss_final.pdf

Ekmekcioglu, E., Stewart T. Worrall & Kondoz, A.M., Low-delay Random View Access in Multi-view Coding using a Bit-rate Adaptive Downsampling Approach. In: *ICME Conference*. 2008, pp. 745–748.

Ekmekcioglu, E., Worrall, S.T. & Kondoz, A.M. (2008). Utilisation of downsampling for arbitrary views in multi-view video coding. *Electronics Letters*. 44 (5). pp. 1–2.

Evans, J. D. (1996). *Straightforward Statistics for the behavioral Sciences*. Pacific Grove, CA: Brooks/Cole Publishing.

Fecker, U. & Kaup, A. H.264/AVC-Compatible Coding of Dynamic of Light Fields using Transposed Picture Ordering. In: *EUSIPCO Conference*. 2005, p. 4.

Fehn, C., Kauff, P., Cho, S., Kwon, H., Hur, N. & Kim, J., Asymmetric Coding of Stereoscopic Video for Transmission Over T-DMB. In: *3DTV Conference*. 2007, pp. 1–4.

Flierl, M., Mavlankar, A. & Girod, B. (2007). Motion and Disparity Compensated Coding for Multiview Video. *IEEE Transactions on Circuits and Systems for Video Technology*. 17 (11). pp. 1474–1484.

Garbas, J.U., Pesquet-Popescu, B. & Kaup, A. (2011). Methods and Tools for Wavelet-based Scalable Multiview Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*. 21 (2). pp. 113–126.

Garcia, D.C., Dorea, C. & de Queiroz, R.L., Super-resolution for Multiview Images using Depth Information. In: *ICIP Conference.* 2010, pp. 1793–1796.

Ghanbari, M. (1999). Video Coding: an Introduction to Standard Codecs. *Institution of Electrical Engineers*.

Guo, X., Y., L., Gao, W. & Huang, Q., Free Viewpoint Switching in Multiview Video. In: *IEEE International Symposium on Circuits and Systems*. 2005, pp. 3471–3474.

Gurler, C.G. & Tekalp, M. (2013). Peer-to-peer System Design for Adaptive 3D Video Streaming. *IEEE Communications Magazine*. 51 (5). pp. 108–114.

Hanhart, P., Ramzan, N., Baroncini, V. & Ebrahimi, T. (2014). Cross-lab Subjective Evaluation of the MVC+D and 3D-AVC 3D Video Coding Standards. In: *2014 Sixth International Workshop on Quality of Multimedia Experience*. 2014, pp. 183–188.

Hannuksela, M.M., Rusanovskyy, D., Su, W., Chen, L., Li, R., Aflaki, P., Lan, D., Joachimiak, M., Li, H. & Gabbouj, M. (2013). Multiview-video-plus-depth Coding based on the Advanced Video Coding Standard. *IEEE Transactions on Image Processing*. 22 (9). pp. 3449–58.

Hewage, C.T.E.R., Appuhami, H.D., Martini, M.G., Smith, R., Jourdan, I. & Rockall, T., Quality Evaluation of Asymmetric Compression for 3D Surgery Video. In: *IEEE 15th International Conference on e-Health Networking, Applications and Services*. 2013, pp. 680–684.

Hussein, H.S., El-Khamy, M. & El-Sharkawy, M. (2013). Blind Configuration of Multi-view Video Coder Prediction Structure. *IEEE Transactions on Consumer Electronics*. 59 (1). pp. 191–199.

ISO/IEC MPEG & ITU-T VCEG (2008). Draft Reference Software for MVC, Available from: https://mvclab.googlecode.com/files/SoftwareManual.doc

ITU-T (2004). Objective Perceptual Video Quality Measurement Techniques for Digital Cable Television in the Presence of a Full Reference Recommendation. Available from:

https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-J.144-200403-I!!PDF-E&type=items

ITU-T (2008). Objective Perceptual Multimedia Video Quality Measurement in the Presence of a Full Reference, Available from: www.ietf.org/mail-archive/web/rmcat/current/pdfliwye67a2T.pdf

Jain, A., Bal, C., Robinson, A., Macleod, D. & Q. Nguyen, T., Temporal Aspects of Binocular Suppression in 3D Video. In: *Sixth international workshop on video processing and quality metrics for consumer electronics*. 2012, p. 6.

Jain, A.K., Robinson, A.E. & Nguyen, T.Q. (2014). Comparing Perceived Quality and Fatigue for Two Methods of Mixed Resolution Stereoscopic Coding. *IEEE Transactions on Circuits and Systems for Video Technology*. 24 (3). pp. 418–429.

Jeon, Y., Sung, J. & Jeon, B., Analysis of Efficient Coding Tools for Multi-view and 3D. In: *IEEE 13th International Symposium on Consumer Electronics*. 2009, pp. 99–102.

Jung, S.-H., Park, W.-J. & Kim, T.-Y., Fast Reference Frame Selection with Adaptive Motion Search using RD Cost. In: *Spring Congress on Engineering and Technology*. 2012, pp. 1–4.

Kalva, H. & Furht, B., Hypercube Based Inter View Prediction for Multi-View Video Coding. In: *Workshop on Immersive Communication and Broadcast Systems*. 2005, pp. 4–7.

Kauff, P., Atzpadin, N., Fehn, C., Muller, M., Schreer, O., Smolic, A. & Tanger, R. (2007). Depth Map Creation and Image-based Rendering for Advanced 3DTV Services Providing Interoperability and Scalability. *Signal Processing: Image Communication*. 22 (2). pp. 217–234.

Kaup, A. & Fecker, U., Analysis of Multi-Reference Block Matching for Multi-View Video Coding. In: *7th Workshop Digital Broadcasting*. 2006, pp. 33–39.

Keimel, C., Deipold, K. & Sarkis, M., Improving the Visual Quality of AVC / H.264 by Combining it with Content Adaptive Depth Map Compression. In: *PCS Conference*. 2010, pp. 494–497.

Khattak, S., Hamzaoui, R., Ahmad, S. & Frossard, P. (2013). Fast encoding techniques for Multiview Video Coding. *Signal Processing: Image Communication*. 28 (6). pp. 569–580.

Kim, S.Y., Cha, J., Lee, S.H., Ryu, J. & Ho, Y.S., 3DTV System using Depth Image-Based Video in the MPEG-4 Multimedia Framework. In: *Proceedings of 3DTV Conference*. 2007, pp. 6–9.

Kimata, H., Kitahara, M., Kamikura, K. & Yashima, Y., Multi-View Video Coding Using Reference Picture Selection For Free-Viewpoint Video Communication. In: *PCS Conference*. 2004, pp. 4–7.

Kitahara, M., Kimata, H., Shimizu, S., Kamikura, K. & Yashima, Y., Multi-view Video Coding using View Interpolation and Reference Picture Selection. In: *ICME Conference*. 2006, pp. 97–100.

Krutz, A. (2010). From Sprites to Global Motion Temporal Filtering, thesis, p.210, Available from: https://opus4.kobv.de/opus4-tuberlin/files/2612/krutz_andreas.pdf

Krutz, A., Drose, M., Kunter, M., Mandal, M., Frater, M. & Sikora, T., Low Bit-Rate Object-based Multi-view Video Coding using MVC. In: *3DTV Conference*. 2007, pp. 1–4.

Kwon, D. & Driessen, P., Efficient and Fast Predictive Block Motion Estimation for Low Bit Rate Video Coding. In: *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*. 2001, pp. 477–480.

L., P. & P., A., Non-uniform Asymmetric Coding of Stereo Images. In: *2nd Romeo Workshop*. 2013, p. 6.

Lee, C., Lee, J., Lee, S., Lee, K., Choi, H., Seo, G. & Park, J. (2011). Full Reference Video Quality Assessment for Multimedia Applications. In: *Recent Researches in Communications, Automation, Signal Processing, Nanotechnology, Astronomy and Nuclear Physics*. pp. 206–209.

Lee, C., Oh, K. & Ho, Y., View Interpolation Prediction for Multi-view Video Coding. In: *PCS Conference*. 2007, pp. 1–4.

Lee, S., Wey, H.-C., Park, D.-S. & Kim, C.-Y., Multi-view Prediction Structure for Free Viewpoint Video. *ICIP Conference*. 2010, pp. 3409–3412.

Lee, Y.-L. (2013). Encoding and Decoding Multi-view Video while Accommodating Absent or Unreliable Camera Parameters. US Patent no 20120269257. p. 18, available from: http://www.google.com.na/patents/US20120269257

Li, S.L.S., Yu, M.Y.M., Jiang, G.J.G., Choi, T.-Y.C.T.-Y. & Kim, Y.-D.K.Y.-D., Approaches to H.264-based Stereoscopic Video Coding. In: *Third International Conference on Image and Graphics (ICIG'04)*. 2004, pp. 2–5.

Li, W. & Ding, G., Panorama-based Multi-view Video Coding. In: *International Conference on Audio, Language and Image Processing*. 2008, pp. 375–379.

Lin, C.-H., Liu, J.-C. & Liao, C.-W., Energy Analysis of Multimedia Video Decoding on Mobile Handheld Devices. In: *TENCON*. 2007, p. 6.

Liu, Y., Huang, Q., Ma, S., Zhao, D., Gao, W., Ci, S. & Tang, H. (2011). A Novel Rate Control Technique for Multiview Video Plus Depth based 3D video Coding. *IEEE Transactions on Broadcasting*. 57 (2). pp. 562–571.

Liu, Y., Huang, Q., Zhao, D. & Gao, W., Low-delay View Random Access for Multi-view Video Coding. In: *IEEE International Symposium on Circuits and Systems*. 2007, pp. 997–1000.

Lu, F., An, P., Zhang, Z. & Shen, L., Multi-view Video Coding Based on Sequence Correlation. In: *International Conference on Audio Language and Image Processing*. 2010, pp. 1227–1232.

Lv, X. (2013). Multi- view Video Coding Scheme Based upon Enhanced Random Access Capacity. *International Journal of Computer Sciences*. 10 (1). pp. 285–289.

Marpe, D., Wiegand, T. & Gordon, S., H.264/MPEG4-AVC Fidelity Range Extensions: Tools, Profiles, Performance, and Application Areas. In: *ICIP Conference*. 2005, p.4.

Marpe, D., Wiegand, T. & Sullivan, G.J. (2006). The H.264/MPEG4 Advanced Video Coding Standard and its Applications. *IEEE Communications Magazine*. 44 (8). pp. 134–143.

Merkle, P., Muller, K., Smolic, A. & Wiegand, T., Efficient Compression of Multi-View Video Exploiting Inter-View Dependencies Based on H.264/MPEG4-AVC. In: *IEEE International Conference on Multimedia and Expo*. 2006, pp. 1717–1720.

Merkle, P., Smolic, A., Muller, K. & Wiegand, T. (2007). Efficient Prediction Structures for Multiview Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*. 17 (11). pp. 1461–1473.

Merkle, P., Smolic, A., Müller, K. & Wiegand, T., Coding Efficiency and Complexity Analysis of MVC Prediction Structures. In: *European Signal Processing Conference*. 2007, pp. 5–9.

Merkle, P., Wang, Y., Muller, K., Smolic, A. & Wiegand, T., Video Plus Depth Compression for Mobile 3D Services. In: *3DTV Conference.* 2009, pp. 1–4.

Miao, G., Himayat, N., Li, Y. & Swami, A. (2009). Cross-Layer Optimization for Energy-Efficient Wireless Communications: A Survey. *Wireless Communications and Mobile Computing*. 9. pp. 529–542.

Mignone, V., Vazquez-Castro, M. a. & Stockhammer, T. (2011). The Future of Satellite TV: The Wide Range of Applications of the DVB-S2 Standard and Perspectives. *Proceedings of the IEEE*. 99 (11). pp. 1905–1921.

Minoli, D. (2011). 3D Television: (3DTV) Technology, Systems, and Deployment: Rolling out the Infrastructure for Next-Generation Entertainment. *CRC Press*, p. 302.

Morvan, Y., Farin, D. & De With, P. (2008). System Architecture for Free-viewpoint Video and 3D-TV. *IEEE Transactions on Consumer Electronics*. 54 (2). pp. 925–932.

Müller, K., Schwarz, H., Marpe, D., Bartnik, C., Bosse, S., Brust, H., Hinz, T., Lakshman, H., Merkle, P., Rhee, F.H., Tech, G., Winken, M. & Wiegand, T. (2013). 3D High-efficiency Video Coding for Multi-view Video and Depth data. *IEEE Transactions on Image Processing*. 22 (9). pp. 3366–78.

Najafi, S. (2012). Single and Multi-view Video Super-resolution Single and Multi-view Video Super-resolution, p.89, Available from: https://macsphere.mcmaster.ca/bitstream/11375/12383/1/fulltext.pdf

Narasak, B., Werapon, C., Kosin, C. & Yo-Sung, H. (2008). Efficient Multiview Video Coding by Object Segmentation. p. 4, Available from: http://icserv.gist.ac.kr/mis/publications/data/2009/narasak.pdf

Nukhet, O. & Tunali, T. (2005). A Survey on the H.264/AVC Standard. *Turkish J. of Electrical Engineering*. 13 (3). pp. 287–302.

Oh, K.J. & Ho, Y.S., Multi-view Video Coding based on the Lattice-like Pyramid GOP Structure. In: *PCS Conference*, 2007, pp. 1–6.

Oh, K.J., Yea, S. & Ho, Y.S., Hole Filling Method Using Depth Based In-painting for View Synthesis in Free Viewpoint Television and 3-D video. In: *PCS Conference.* 2009, p. 6.

Ohm, J., Stereo / Multiview Video Encoding Using the MPEG Family of Standards. In: *Proc. SPIE Conf. Stereoscopic Displays Virtual Reality Syst*. 1999, pp. 1–12.

Oka, S., Endo, T., Fuji, T. & Tanimoto, M. (2004). Dynamic Ray-space Coding using Multi-directional Picture, *IEICE Technical Report,* 104(493), p. 6.

Ostermann, J., Bormans, J., List, P., Marpe, D., Narroschke, M., Pereira, F., Stockhammer, T. & Wedi, T. (2004). Video coding with H.264/AVC: Tools, performance, and complexity. *IEEE Circuits and Systems Magazine*. 4. pp. 7–28.

Ozbek, N. & Murat Tekalp, A.. Scalable Multi-view Video Coding for Interactive 3DTV. In: *ICME Conference*. 2006, pp. 213–216.

Ozbek, N. & Tekalp, a. M., Unequal Inter-view Rate Allocation using Scalable Stereo Video Coding and an Objective Stereo Video Quality Measure. In: *IEEE International Conference on Multimedia and Expo*. 2008, pp. 1113–1116.

Palaniappan, R. & Nikil, J., Subjective Quality in 3DTV : Effects of Unequal Bit Allocation to Left and Right Views. In: *6th International Workshop on Video Processing and Quality Metrics for Consumer Electronics*. 2012, pp. 87–92.

Park, P.-K., Oh, K.-J. & Ho, Y.-S. (2008). Efficient View-temporal Prediction Structures for Multi-view Video Coding. *Electronics Letters*. 44 (2). pp. 5–6.

Paul, M. & Sorwar, G., Encoding and Decoding Techniques for Medical Video Signal Transmission and Viewing. In: *6th IEEE/ACIS International Conference on Computer and Information Science*. 2007, pp. 750–756.

Pedro López Velasco (2012). Video Quality Assessment, Video Compression, Available from: http://www.intechopen.com/books/video- compression/video-quality-assessment

Perkins, M.G. (1992). Data Compression of Stereopairs. *IEEE Transactions on Communications*. 40 (4). pp. 684–696.

Pinto, L. & Assuncao, P., Asymmetric 3D Video Coding using Regions of Perceptual Relevance. In: *IC3D Conference*. 2012, p. 6.

Pourazad, M.T., Nasiopoulos, P. & Ward, R.K.. A New Prediction Structure for Multiview Video Coding. In: *DSP Conference*. 2009, pp. 1–5.

Quan, J., Hannuksela, M.M. & Li, H., Asymmetric Spatial Scalability in Stereoscopic Video Coding. In: *3DTV Conference*. 2011, pp. 1–4.

Richardson, I.E. (2010). The H.264 Advanced Video Compression Standard. *2nd Ed. John Wiley & Sons, Ltd*.

Sampaio, F., Zatt, B., Shafique, M., Agostini, L., Bampi, S. & Henkel, J., Energy-Efficient Memory Hierarchy for Motion and Disparity Estimation in Multiview Video Coding. In: *Design, Automation & Test in Europe Conference & Exhibition*. 2013, pp. 665–670.

Sansli, D.B., Ugur, K., Hannuksela, M.. & Gabbouj, M. (2014). Inter view Motion Vector Prediction in Multiview HEVC. In: *2014 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*. 2014, pp. 1–4.

Savas, S.S., Gurler, C.G. & Tekalp, A.M., Quality of Experience of Multi-View Video for IP Delivery. In: *19th International Packet Video Workshop*. 2012, p. 6.

Saygili, G., Gurler, C.G. & Tekalp, A.M. (2011). Evaluation of Asymmetric Stereo Video Coding and Rate Scaling for Adaptive 3D Video Streaming. *IEEE Transactions on Broadcasting*. 57 (2). pp. 593–601.

Saygili, G., Gürler, C.G. & Tekalp, A.M., Quality Assessment of Asymmetric Stereo Video Coding. In: *ICIP Conference*. 2010, pp. 3–6.

Schwarz, H., Marpe, D. & Wiegand, T. (2007). Overview of the Scalable Video Coding Extension of the H.264/AVC Standard. *IEEE Transactions on Circuits and Systems for Video Technology*. 17 (9). pp. 1103–1120.

Schwarz, H., Marpe, D. & Wiegand, T., Analysis of Hierarchical B Pictures and MCTF. In: *IEEE International Conference on Multimedia and Expo*. 2006, pp. 1929–1932.

Seungwook, H. & Yang, Y. (2011). Dynamic reference frame reordering for frame sequential stereoscopic video encoding. *Patent*, pp. 1–9.

Shafique, M., Zatt, B., Bampi, S. & Henkel, J., Power-aware Complexity-scalable Multiview Video Coding for Mobile Devices. In: *PCS Conference.* 2010, pp. 350–353.

Shao, F., Jiang, G., Yu, M., Chen, K., Ho, Y. & Member, S. (2012). Asymmetric Coding of Multi-View Video Plus Depth Based 3-D Video for View Rendering. *IEEE Transactions on Multimedia*. 14 (1). pp. 157–167.

Shao, L., Kirenko, I., Leitao, A. & Mydlowski, P., Motion-Compensated Techniques for Enhancement of Low-Quality Compressed Videos. In: *ICASSP Conference*. 2009, pp. 1349–1352.

Sheikh Akbari, A., Canagarajah, N., Redmill, D., Bull, D. & Agrafiotis, D., A Novel H.264/AVC Based Multi-View Video Coding Scheme. In: *3DTV Conference*. 2007, pp. 1–4.

Shen, L., Liu, Z., Yan, T., Zhang, Z. & An, P. (2010). View-Adaptive Motion Estimation and Disparity Estimation for Low Complexity Multiview Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*. 20 (6). pp. 925–930.

Shen, L., Liu, Z., Zhang, Z. & Wang, G. (2007). An Adaptive and Fast Multiframe Selection Algorithm for H.264 Video Coding. *IEEE Signal Processing Letters*. 14 (11). pp. 836–839.

Shen, L., Liu, Z., Zhang, Z. & Wang, G., Video Nature Considerations for Multi-frame Selection Algorithm in H.264. In: *IEEE/ACS International Conference on Computer Systems and Applications*. 2007, pp. 708–711.

Smirnov, S., Gotchev, A., Sumeet, S., Gerhard, T. & Heribert, B., 3D Video Processing Algorithms – Part I*, Mobile 3DTV*. 2010, p.42.

Smolić, A. & Kauff, P. (2005). Interactive 3-D Video Representation and Coding Technologies. *Proceedings of the IEEE*. 93 (1). pp. 98–110.

Smolic, A. (2011). 3D Video and Free Viewpoint Video - from Capture to Display. *Pattern Recognition*. 44 (9). pp. 1958–1968.

Smolic, A., Mueller, K., Merkle, P., Fehn, C., Kauff, P., Eisert, P. & Wiegand, T., 3D Video and Free Viewpoint Video - Technologies, Applications and MPEG Standard. In: *ICME Conference*. 2006, pp. 2161–2164.

Smolic, A., Mueller, K., Stefanoski, N., Ostermann, J., Gotchev, A., Akar, G.B., Triantafyllidis, G. & Alper, K. (2007). Coding Algorithms for 3DTV — A Survey. *IEEE Transactions on Circuits and Systems for Video Technology*. 17 (11). pp. 1606–1621.

Stelmach, L., Tam, W.J., Meegan, D., Vincent, A. & H, H. (2000). Stereo Image Quality : Effects of Mixed Spatio-Temporal Resolution. *IEEE Transactions on Circuits and Systems for Video Technology*. 10 (2). pp. 188–193.

Stoykova, E., Alatan, A.A., Benzie, P., Grammalidis, N., Malassiotis, S., Ostermann, J., Piekh, S., Sainov, V. & Theobalt, C. (2007). 3-D Time-Varying Scene Capture Technologies — A Survey. *IEEE Transactions on Circuits and Systems for Video Technology*. 17 (11). pp. 1568–1586.

Strohmeier, D. & Tech, G., On Comparing Different Codec Profiles of Coding Methods for Mobile 3D Television and Video. In: *International Conference on 3D Systems and Applications*. 2010, pp. 1–4.

Su, Y., Vetro, A. & Smolic, A. (2006). Common Test Conditions for Multiview Video Coding, *report*, Available from: https://comment-jmvc.googlecode.com/files/JVT-T207.doc

Sühring, K. (2011). JM reference software version 18.0. *Fraunhofer HHI*. Available from: http://iphome.hhi.de/suehring/tml/download/old_jm/.

Sullivan, G.J., Ohm, J., Han, W. & Wiegand, T. (2012). Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transactions on Circuits and Systems for Video Technology*. 22 (12). pp. 1649–1668.

Sullivan, G.J., Topiwala, P. & Luthra, A., The H.264/AVC Advanced Video Coding Standard : Overview and Introduction to the Fidelity Range Extensions. In: *SPIE Conference on Applications of Digital Image Processing XXVII*. 2004, pp. 1–22.

Tanimoto, M. (2012). FTV: Free-viewpoint Television. *Signal Processing: Image Communication*. 27 (6). pp. 555–570.

Tanimoto, M., Overview of FTV (Free-Viewpoint Television). In: *ICME Conference*. 2009, pp. 1552–1553.

Tech, G. (2012). 3D-HEVC Test Model 1. *Proceedings of the Meeting of Joint Collaborative Team on 3D Video Coding*. 2012, p. 83

Tech, G., Brust, H., Müller, K., Aksay, A. & Bugdayci, D. (2009). Development and Optimization of Coding Algorithms for Mobile 3DTV. Available from: http://sp.cs.tut.fi/mobile3dtv/results/tech/D2.5_Mobile3DTV_v1.0.pdf.

Tech, G., Chen, Y., Muller, K., Ohm, J.-R., Vetro, A. & Wang, Y.-K. (2015). Overview of the Multiview and 3D Extensions of High Efficiency Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*. (99). p.14

Tech, G., Smolic, A., Brust, H., Merkle, P., Dix, K., Wang, Y., Muller, K. & Wiegand, T., Optimization and Comparison of Coding Algorithms for Mobile 3DTV. *3DTV Conference*. 2009, pp. 1–4.

Tech, V.S.M. & Babu, K.A. (2011). Low Bit-Rate Image Compression using Adaptive Down-Sampling Technique. *IJCTA*. 2 (5). pp. 1679–1689.

Tian, D., Lai, P.-L., Lopez, P. & Gomila, C. (2009). View Synthesis Techniques for 3D Video. *Proceedings of SPIE*. 7443 (609). p. 11.

Tsung, P.-K., Ding, L.-F., Chen, W.-Y., Chien, S.-Y., Ch, T.-C. & Chen, L.-G., System Bandwidth Analysis of Multiview Video Coding with Precedence Constraint. In: *IEEE International Symposium on Circuits and Systems*. 2007, pp. 1001–1004.

Ugur, K., Liu, H., Lainema, J., Gabbouj, M. & Li, H., Parallel Encoding - Decoding Operation for Multiview Video Coding with High Coding Efficiency. In: *3DTV Conference*. 2007, pp. 1–4.

Uslubas, S., Maani, E. & Katsaggelos, A.K. (2010). A Resolution Adaptive Video Compression System. *Intelligent Multimedia Communication: Techniques and Applications*. 280. pp. 167–194.

Vetro, a, Wiegand, T. & Sullivan, G.J. (2011). Overview of the Stereo and Multiview Video Coding Extensions of the H.264/MPEG-4 AVC Standard. *Proceedings of the IEEE*. 99 (4). pp. 626–642.

Vetro, A. (2010). Representation and Coding Formats for Stereo and Multiview Video. *Intelligent Multimedia Communication: Techniques and Applications*. 280. pp. 51–73.

Wang, Z., Bovik, A.C., Sheikh, H.R. & Simoncelli, E.P. (2004). Image Quality Assessment : From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*. 13 (4). pp. 1–14.

Wei, X. (2007). Efficient Multi-view Video Coding Scheme Based on Dynamic Video Object Segmentation, *thesis*, p. 152, Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128.7631&rep=rep1&type=pdf

Wong, K.-M., Po, L.-M., Cheung, K.-W., Ng, K.-H. & Xu, X.. Stretching, Compression and Shearing Disparity Compensated Prediction Techniques for Stereo and Multiview Video Coding. In: *ICASSP Conference*. 2011, pp. 841–844.

Xu, X. & He, Y., Fast Disparity Motion Estimation in MVC Based on Range Prediction. In: *ICIP Conference*. 2008, pp. 2000–2003.

Yamamoto, K., Kitahara, M., Kimata, H., Yendo, T., Fujii, T., Tanimoto, M., Member, S., Shimizu, S., Kamikura, K. & Paper, I. (2007). Multiview Video Coding Using View Interpolation and Color Correction. *IEEE Transactions on Circuits and Systems for Video Technology*. 17 (11). pp. 1436–1449.

Yang, H., Yu, M. & Jiang, G., Decoding and Up-sampling Optimization for Asymmetric Coding of Mobile 3DTV. In: *TENCON - IEEE Region 10 Conference*. 2009, pp. 1–4.

Yang, P. & He, Y., Diagonal Interview Prediction for Multiview Video Coding. In: *PCS Conference*. 2007, p. 4.

Yang, W., Lu, Y., Wu, F., Cai, J., Ngan, K.N. & Li, S. (2006). 4-D wavelet-based multiview video coding. *IEEE Transactions on Circuits and Systems for Video Technology*. 16 (11). pp. 1385–1395.

Yang, Z. & Goodwin, N., A Bandwidth Management Framework for Wireless Camera Array. In: *NOSSDAV*. 2005, p. 6.

Yea, S. & Vetro, A. (2009). View synthesis prediction for multiview video coding. *Signal Processing: Image Communication*. 24 (1-2). pp. 89–100.

Yebin, L., Qionghai, D. & Wenli, X., A Real Time Interactive Dynamic Light Field Transmission System. In: *ICME Conference*. 2006, pp. 2173–2176.

Yoon, H. & Kim, M. (2012). Temporal Prediction Structure for Multi-view Video Coding. *Journal of Korea Multimedia Society*. 15 (9). pp. 1093–1101.

Yu, H. & Winkler, S., Image Complexity and Spatial Information. In: *5th International Workshop on Quality of Multimedia Experience*. 2013, pp. 12–17.

Yu, M., Yang, H., Fu, S., Li, F., Fu, R. & Jiang, G., New Sampling Strategy in Asymmetric Stereoscopic Video Coding for Mobile Devices. In: *International Conference on E-Product E-Service and E-Entertainment*. 2010, pp. 1–4.

Yuan, H., Kwong, S., Wang, X., Gao, W. & Zhang, Y, (2015). Rate Distortion Optimized Inter-View Frame Level Bit Allocation Method for MV-HEVC. *IEEE transactions on multimedia*. 17 (12). pp. 2134–2146.

Yu-wen, H., Bing-yu, H., Shao-yi, C., Shyh-yih, M. & Liang-gee, C. (2006). Analysis and Complexity Reduction of Multiple Reference Frames Motion Estimation in H.264/AVC. *transaction on Circuits and Systems for Video Technology*. 16 (4). pp. 507–522.

Zainaldin, A., Lambadaris, I. & Nandy, B., Adaptive Rate Control Low Bit-Rate Video Transmission over Wireless Zigbee Networks. In: *IEEE International Conference on Communications*. 2008, pp. 52–58.

Zhang, L., Kang, J., Zhao, X., Chen, Y. & Joshi, R., Neighboring Block Based Disparity Vector Derivation for 3D-AVC. In: *Visual Communications and Image Processing (VCIP)*. 2013, pp. 1–6.

Zhang, Y., Jiang, G., Yi, W., Yu, M., Jiang, Z. & Kim, Y.D., An Approach to Multi-modal Multi-view Video Coding. In: *ICSP Conference*. 2006, pp. 2–5.

Zhang, Y., Jiang, G.Y., Yu, M. & Ho, Y.S. (2009). Adaptive Multiview Video Coding Scheme Based on Spatiotemporal Correlation Analyses. *ETRI Journal*. 31 (2). pp. 151–161.

Zhang, Y., Kwong, S., Jiang, G. & Wang, H. (2011). Efficient Multi-Reference Frame Selection Algorithm for Hierarchical B Pictures in Multiview Video Coding. *IEEE Transactions on Broadcasting*. 57 (1). pp. 15–23.

Zhang, Y., Mei, Y. & Gangyi, J. (2008). Evaluation of Typical Prediction Structures for Multi-view Video Coding. *ISAST Transactions on Electronics and Signal Processing*. 2. pp. 7–15.

Zitnick, C.L., Kang, S.B., Uyttendaele, M., Winder, S. & Szeliski, R. (2004). High-Quality Video View Interpolation using a Layered Representation. *ACM Transactions on Graphics*. 23 (3). pp. 600–608.

PUBLICATIONS

- Sheikh Akbari, A., Said, H. & Moniri, M. Effect of Inter-camera Angles on the Performance of an H.264/AVC Based Multi-view Video Codec. In: *Picture Coding Symposium*. May 2012, IEEE, pp. 109–112.

- Said, H. & Sheikh Akbari, A. H.264/AVC Based Multi-view Video Codec using the Statistics of Block Matching. In: *ELMAR, 55th International Symposium*. 2013, pp. 97–100.

- Said, H., Sheikh Akbari, A. & Moniri, M. An Adaptive Reference Frame Re-Ordering Algorithm for H.264/AVC Based Multi-View Video Codec. In: *EUSIPCO*. 2013, pp. 1–5.

Effect of Inter-Camera Angles on the Performance of an H.264/AVC based Multi-View Video Codec

[1]Akbar Sheikh Akbari, [2]Hany Said and [3]Mansour Moniri
Faculty of Computing, Engineering and Technology
Staffordshire University, Beaconside
Stafford, UK
Email: {[1]a.s.akbari, [2]h.h.said, [3]m.moniri}@staffs.ac.uk

Abstract— **This paper investigates the effect of inter-camera angles on the performance of an H.264/AVC based multi-view video codec. To achieve this, the H.264/AVC software has been modified to support multi-view video coding using its multi-frame reference property. Results were generated using a wide baseline convergent multi-view video data set: Breakdancers. To generate a set of three synchronized multi-view videos from the same scene with different inter-camera angles, all possible three camera combinations are generated and classified according to their inter-camera angles. The resulting set of multi-view videos are coded using H.264/AVC based multi-view and simulcast video codecs at different bitrates. Results demonstrate that the multi-view video codec gives superior coding performance up to 1.2dB compared to that of simulcast coding scheme at low inter-camera angles and it deteriorates as the inter camera angles increase. Finally, a range of inter-camera angles for best use of either multi-view or simulcast coding is determined.**

Keywords-multi-view video codec; H.264/AVC; inter-camera angles.

I. INTRODUCTION

3D and free viewpoint video are new types of natural video media that expand the user's sensation far beyond what is offered by traditional media. The first offers 3D depth impression of the observed scenery, while the second allows for interactive selection of viewpoint and direction within a certain operating range as known from computer graphics applications [1]. However, the price for utilizing the natural video media enormously increases the amount of data to be stored or transmitted. A multi-view imaging environment consists of an array of cameras, which image the world scene from different positions and viewing angles. As the number of camera views increases, the size of the dataset increases linearly. Since all the cameras capture the same world scene, there is a colossal amount of correlation within the multi-view sequences. To achieve a good trade-off between scene quality and bitrate, disparity and motion among all the frames have to be efficiently exploited. Predictive coding is one of the techniques that is widely used to perform both disparity and motion compensation [2]. In predictive coding, the previously decoded frames are used as references to predict the current frame. Disparity compensation view prediction exploits correlation among the views using motion compensation prediction concept [3]. H.264/AVC is the latest development in monoscopic video coding schemes that supports multiple reference frame motion compensation prediction. This feature of H.264/AVC has made it efficient for coding multi-view video sequences. Several multi-view video codecs using H.264/AVC have already been reported in the literature [4-6]. A H.264/AVC based multi-view (MV) video coding scheme was reported in [7]. This codec supports both view and temporal scalability, while it offers better subjective and objective (up to 0.5 dBs) quality for certain sequences compared to simulcast coding. Another H.264/AVC based MV video codec was proposed in [8]. This codec first generates a synthesized prediction frame from the decoded frames of the neighbouring views using camera parameters and depth-map information. The synthesized prediction frame is then used for disparity compensation view prediction. A coding gain of up to 2 dBs higher than simulcast coding when coding ballroom MV-sequences was reported. Markle et al. also employed an extended version of the H.264/AVC software to compress both multi-view color and depth information [9]. They compressed the color and depth data of the Breakdancer multi-view sequences using their proposed codec and simulcast coding scheme and reported objective gain of about 0.5 dBs higher than that of simulcast coding for coding both color and depth data. Another MV video coding scheme based on a lattice-like pyramid GOP structure was suggested by Oh and Ho in [10]. In their proposed codec the number of intra-frames are almost reduced to half compared to that of anchor coding by using RB frames (an RB-frame is a frame that predict solely from frames of two neighbouring views.), where RB-frames play the role of intra-frames in coding processes of their view sequence. The coding performance of up to 1dBs greater than anchor coding was reported. A H.264/AVC based asymmetric video codec was presented by Chen et al. in [11]. Their proposed codec applies a regionally adaptive filtering algorithm to generate a prediction for the low-resolution view from the high resolution view. The codec targets stereoscopic video applications with a bandwidth slightly higher than that of having two mono-view video communications with comparable subjective video quality. Results show that Chen et al.'s codec provides about 8% bit-rate saving on average, and 27% bit-rate saving at

most, which is equivalent to more than 0.7 dB luma Peak Signal-to-Noise (PSNR) gain for the low-resolution view. However, the effect of inter camera angles, especially for wide baseline camera setups, on coding performance of the multi-view video codecs has not been reported in the literature. M. Abdoli et al. [4] investigated the impact of view spacing on the total contribution of cross view block prediction in multi-view video codecs. They showed that the contribution of cross view block prediction has an indirect relationship with view spacing and is significantly reduced as the view spacing is increased. However, the effect of inter camera angles, especially for wide baseline camera setups, on coding performance of the multi-view video codecs has not been reported in the literature.

In this paper, the effect of inter camera angle on the performance of the H.264/AVC based multi-view video codec is investigated. From experimental results, inter-camera angle threshold will be determined for the best use of multi-view video codecs and simulcast video coding. The rest of the paper is arranged as follows; Dataset preparation will be presented in Section II. In Section III, H.264/AVC based multi-view video codec is introduced. Experimental results are given in Section IV. Finally Section V concludes the paper.

II. DATASET PREPARATION

In order to investigate the effect of inter-camera angles on the performance of the H.264/AVC based multi-view video codec, a number of multi-view video datasets must be first generated. This entails capturing the same scene using three or more convergent cameras that have different inter-camera angles. In this research, Breakdancers multi-view sequence, generated by Microsoft laboratories using eight synchronized PtGrey color cameras, has been chosen [12]. These eight cameras were convergent on the circumference of a horizontal arc spanning 30 degrees. Based on the camera setup presented in [12], it can be calculated that the cameras were located at a distance of 6 meters away from the scene. The sequence contains eight videos with one hundred frames and a resolution of 1024×768. These video sequences were captured by converging cameras from different viewing angles at 15 frames per second. Each camera has a 30 degree field of view with an 8 mm lens and was calibrated using a 36" ×36" calibration pattern mounted on a flat plate, with Zhang's calibration techniques [13]. As this research targets the application of multi-view video transmission over low bitrate channels, the resolution of the input dataset is reduced to the CIF size. To achieve this, all the frames of the dataset are processed as follows: a) each input frame is filtered using a 3×3 FIR Blackman low-pass filter (the coefficients of this filter are tabulated in Table I); b) each filtered frame is then down-sampled by a factor 2 both horizontally and vertically; c) to maintain the external camera parameters unchanged and also to keep the most of foreground contents of the multi-view data set, each resulting down-sampled frame is then cropped from point $(P_x, P_y) = (130, 68)$ onward to make a CIF size RGB frame for it; d) the resulting CIF size RGB frames are finally converted to YUV in full color sampling 4:4:4 format.

To generate a number of multi-view video datasets with different inter-camera angles from the Breakdancers dataset, all

TABLE I. BLACKMAN FIR FILTER COEFFICIENTS.

0.0381	0.1051	0.0381
0.1051	0.4273	0.1051
0.0381	0.1051	0.0381

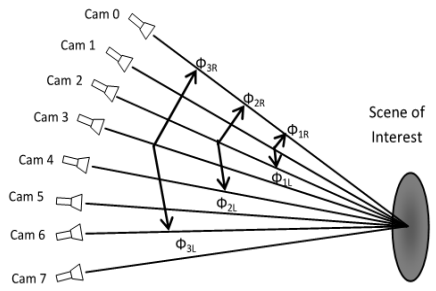


Fig. 1. Different inter-camera angles for camera setups of multi-view Breakdancers' data set.

TABLE II. CAMERA SEPARATION ANGLES FOR BREAKDANCERS MULTI-VIEW SEQUENCES WITH RESPECT TO C4.

Camera number	Inter-camera angle (deg)
0	-15.8
1	-12.6
2	-9.25
3	-4.63
4	0
5	+2.69
6	+7.52
7	+10.76

combinations of the three multi-view video sequences with different inter-camera angles, Φ_1 , Φ_2 and Φ_3 , as shown in Figure 1, were created. In this figure, optical line of each camera is shown as a line from the camera to the centre of the scene. The inter-camera angles Φ_{iR} and Φ_{iL} represent the angle between the center camera and the camera to the right or left, respectively, and i represents the general angle multiplier for inter-camera angles. Table II lists the inter-camera angles for Breakdancers. The inter-camera angles were calculated by extracting the panning angles from the camera rotation matrices provided in [13], with respect to camera C_4. Having calculated these angles in reference to camera C_4, it is possible

TABLE III. INTER-CAMERA ANGLES FOR MULTI-VIEW VIDEO SEQUENCES.

Φ_i	Φ_1					
MV sequence	012	123	234	345	456	567
Φ_{iR}	3.2	3.35	4.62	4.63	2.69	4.83
Φ_{iL}	3.35	4.62	4.63	2.69	4.83	3.24
Φ_i	Φ_2				Φ_3	
MV sequence	024	135	246	357	036	147
Φ_{iR}	6.55	7.97	9.25	7.32	11.17	12.15
Φ_{iL}	9.25	7.32	7.52	8.07	12.15	10.76

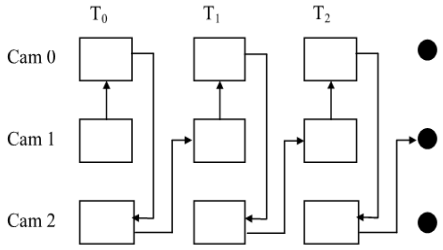


Fig. 2. Interleaving Multi-view videos to generate a single stream video.

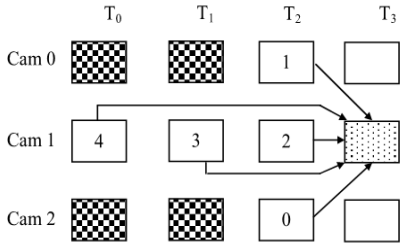


Fig. 3. Temporal and cross-view referencing for the proposed multi-view video codec.

to calculate the inter-camera angle between any pair of cameras used in capturing the video datasets. Table III shows inter-camera angles for all possible combination of the three sets of multi-view video sequences with different inter-camera angles. From this table, it can be seen that there are 6, 4 and 2 multi-view video sequences with different inter-camera angles.

III. H.264/AVC BASED MULTI-VIEW VIDEO CODEC

In order to use the monoscopic H.264/AVC codec to encode multi-view videos, the multi-view sequences must be merged

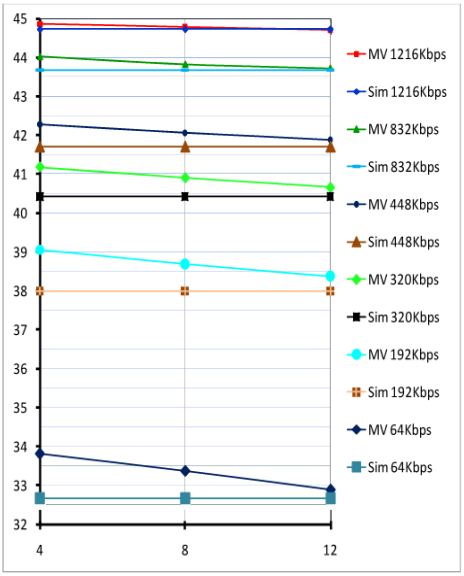


Fig. 4. Average $PSNR_Y$ for Simulcast and multi-view video codec at different bitrates and inter-camera angles.

into one video sequence prior to the encoding. Figure 2 shows how the frames from different cameras have been interleaved to generate a mono video stream. From this figure, it can be seen that interleaving has been performed by taking one frame from the central view and one frame from each of the two neighboring views. Multiple frame referencing enables more than one previously coded P- or B-frames, and also the latest I-frame, to be used for prediction in both P- and B-frames [14]. Figure 3 shows the block diagram of the reference frames being used in the proposed multi-view video codec, where blocks 0 to 4 represent the latest 5 decoded frames in the H.264/AVC frame buffer. Only the two most recently coded frames of the neighboring views and the latest three coded frames of the current view are used to predict the current frame. In this research, for simplicity, a $I \quad P \quad P \quad \cdots$ stream is chosen to generate the experimental results. The decoded Picture Buffer (DPB) of the H.264/AVC has been modified to support the given multi-view architecture.

IV. EXPERIMENTAL RESULTS

The proposed H.264/AVC based multi-view video codec and a simulcast H.264/AVC standard video codec are used to code the selected videos at 15fps, and 10 different bitrates starting at 64kbps with steps of 128kbps. The average PSNR of the decoded luminance component of the videos was calculated at different bitrates and results are shown in Figure 4. From this figure, it can be seen that the multi-view video codec gives superior performance, by up to 1.4 dBs compared to that of simulcast coding for almost all bitrates. From this figure, it can be noticed that the coding performance of the multi-view video codec deteriorates as the inter-camera angle increases for all

bitrates. The coding performance of the multi-view video codec is almost the same as simulcast coding scheme when coding videos with inter-camera angle of 12 degrees. This figure also shows that the coding performance of the multi-view video codec decreases as channel bitrates is increased. The coding performance of the multi-view codec dropped almost to a value very closed to that of simulcast coding at high bitrates. It implies that at high bitrates the application of multi-view video codec does not deliver higher coding performance compared to simulcast coding, although it adds huge computational cost to the block matching stage of the codec. In other words, the multi-view video codec produces superior coding performance to that of simulcast coding at lower bitrates and smaller inter-camera angles. Increasing the computational cost for finding the best matches amongst reference frames from both cross views and temporal reference frames is the overhead.

The difference between the coding performance of the multi-view video codec and simulcast codec for coding the Breakdancers sequences are shown in Table IV. From this table it is clear that the multi-view video codec gives superior coding performance (up to 1.4 dBs) to that of simulcast coding at low bitrates and also smaller inter-camera angles. Results in this table also reveal that the multi-view video codec gives superior coding performance, more than almost 0.5 dB, to that of simulcast coding at inter camera angles less than 8 degrees and bitrates below 448 kbps. This is the best range for the use of multi-view video codecs.

TABLE IV. PSNR DIFFERENCE OF MULTI-VIEW VIDEO CODEC AND SIMULCAST CODING FOR DIFFERENT INTER-CAMERA ANGLES AT DIFFERENT BITRATES.

Bitrate Kbps	Δ **PSNR**$_Y$ [$PSNR_{MV(\Phi1)}$ - $PSNR_{Sim}$)]	Δ **PSNR**$_Y$ [$PSNR_{MV(\Phi2)}$ - $PSNR_{Sim}$)]	Δ **PSNR**$_Y$ [$PSNR_{MV(\Phi3)}$ - $PSNR_{Sim}$)]
64	1.156	0.7075	0.2195
192	1.0588	0.699	0.3795
320	1.444	0.489	0.2415
448	0.57825	0.36225	0.18075
576	0.434	0.262	0.1285
704	0.3375	0.1945	0.0745
832	0.34725	0.14875	0.04575
960	0.1973	0.0985	0.0845
1088	0.16317	0.0645	-0.01
1216	0.12992	0.04725	-0.02775

I. CONCLUSIONS

In this paper, the effects of inter camera angle on the performance of the H.264/AVC based multi-view video codec for wide baseline converging camera setup was investigated. The Breakdancers multi-view data set was used to generate a number of multi-view video streams with different inter camera angles. The H.264/AVC software was modified to support the multi-view video coding. Experimental results showed that the multi-view video codec performance decreases as the inter-camera angle increases. Based on the experimental results a range for best use of either multi-view or simulcast codec was determined.

REFERENCES

[1] A. Smolic, and P. Kauff, "Interactive 3-D Video Representation and Coding Technologies", Proceedings of the IEEE, vol. 93, issue 1, pp. 98-110, 2005.

[2] M. Flierl, A. Mavlankar, and B. Girod, "Motion and Disparity Compensated Coding for Multiview Video", IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, issue 11, pp. 1474-1484, November 2007.

[3] K. Muller, P. Merkle, H. Schwarz, T. Hinz, A. Smolic, T. Wieqand and T Oelbaum, "Multi-view video coding based on H.264/AVC using hierarchical B-frames", Proceedings of the Picture Coding Symposium, Beijing, China, April 2006.

[4] M. Abdoli, M. Soryani and A.F. Aminian odarres, "The Impact of view spacing in Multiview video compression efficiency", Seventh International Conference on Information Technology: New Technology, pp. 1314-1315, July 2010.

[5] S. A. Fezza, K. M. Faraoun, and S. Ouddane, "A Comparison of Prediction Structures for Multi-view Video Coding Based on The H.264/AVC Standard", 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA), pp.111-114, 2011.

[6] A. Vetro, T. Weigand, 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, issue 4, pp. 626-642, March 2011.

[7] M. Drose, C. Clemens, and T. Sikora, "Extending Single-View Scalable Video Coding to Multi-View Based on H.264/AVC", IEEE iInternational Conference on Image Processing (ICIP), pp. 29-77-2980, Atlanta, USA, October 2006.

[8] S. Ince, E. Martinian, Sehoon Yea and A. Vetro, "Depth Estimation for View Synthesis in Multiview Video Coding", 3DTV conference, Kos Island, Greece, May 2007.

[9] P. Merkle, A. Smolic, K. Müller and T. Wiegand, "Efficient Compression of Multi-View Depth Data Based on MVC", 3DTV conference, Kos Island, Greece, May 2007.

[10] K.-J. Oh, and Y.-S. Ho, "Multi-view Video Coding based on the Lattice-like Pyramid GOP Structure", Proceedings of the Picture Coding Symposium, Beijing, China, April 2006.

[11] Y. Chen, Y.-K. Wang, M. Gabbouj, and Miska M. Hannuksela, "Regionally Adaptive Filtering for Asymmetric Stereoscopic Video Coding", in Proceedings of the IEEE International Symposium on Circuits and Systems, ISCAS 2009, Taipei, Taiwan, pp. 2585-2588, May 2009,

[12] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder and R. Szeliski, "High-quality video view interpolation using a layered representation" The ACM SIGGRAPH and ACM Transaction on Graphics, Los Angeles, CA, August 2004.

[13] Z. Zhang, "A flexible new technique for camera calibration", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, issue 11, pp. 1330-1334, November 2000.

[14] C. Bilen, A. Aksay, G.B. Akar, "A Multiview video codec based on H.264", IEEE International Conference on Image Processing (ICIP) , pp. 541-544, October 2006.

H.264/AVC Based Multi-view Video Codec using the Statistics of Block Matching

Hany Said, Akbar Sheikh Akbari
Faculty of Computing, Engineering and Sciences,
Staffordshire University, Beaconside, Stafford, UK
h.h.said@staffs.ac.uk

***Abstract* - This paper proposes two reference frame architectures for H.264/AVC based multi-view video codecs. To achieve this, the block matching amongst reference frames of the codec are statistically analyzed. Based on the resulting statistics, two sets of reference frame architectures for best coding performance of the codec are proposed. The coding performance of the codec using the proposed reference frame architectures are assessed against the same codec which uses three different reference frame architectures. The measurements were carried out on four standard multi-view datasets. Results show that the application of the proposed reference frame architectures significantly (up to 2.3 dBs) improves the coding performance of the codec.**

***Keywords* - Video compression, H.264/AVC, multi-view video codec, reference frame ordering**

I. INTRODUCTION

3D-TV and Free View point media applications offer a very high interactive representation of the natural scene and provides more independence to the end user at the cost of high computational complexity and resources. These media applications have opened the research corridors to address and exploit implications involved in improving the technology and its compatibility with available framework resources. The main contention is the enormous amount of data that needs to be processed, stored and transmitted over the available resources [1]. Multi-view imaging environment comprises of an array of cameras to capture the scene from different viewing angles. The greater the number of cameras, the greater the data needed to be processed. Since each camera captures the same scene from different viewing angle, there exists a very high correlation between the views captured from different camera of the array. This correlation is exploited by motion and disparity compensation for efficient compression [2]. H.264/AVC offers motion and disparity prediction and compensation by its multiple frames referencing properties, which makes this codec efficient for multi-view video encoding [3-5].

Many H.264/AVC based Multi-View Video (MVV) codecs have been reported in the literature [6-10]. These codecs use various reference frame architectures to efficiently encode P- and B-frames. A H.264/AVC based MVV-codec, which gives superior coding performance to simulcast coding, was reported in [6]. The proposed codec uses one temporal, one adjacent spatial neighboring frame and two adjacent spatiotemporal-frames to generate a prediction for the current frame. Another H.264/AVC based MVV-codec was proposed in [7]. This codec uses all spatial and three temporal frames to efficiently predict current frame's macroblocks. Application of these reference frames significantly improved the coding gain of the codec compared to the use of all spatial, three temporal and all spatiotemporal reference frames. Bilen *et al.* proposed a MVV-codec based on H.264/AVC codec. Their codec supports three reference frame architectures. Its first architecture contains two temporal, one adjacent spatial and, two-spatiotemporal frames. The second architecture uses two temporal, one adjacent spatial and one spatiotemporal frames and the third architecture includes one temporal, one adjacent spatial and one spatiotemporal frames [8]. Other reference frame architecture for H.264/AVC based multi-view codec was proposed by Fecker *et al.* [9]. They first arranged the pictures in a transposed order and proposed a prediction structure that uses the last $N+1$ reference frames, where N is the number of views. They reported significant coding performance (up to 6 dBs) in compared to simulcast video coding. Said *et al.* [10] were performed two sets of statistical analysis of block matching amongst reference frames of a stereoscopic video codec. They showed that there is a relationship between the order of reference frames and coding performance of the codec. They also reported an improvement in coding performance of the codec when it employs reference frames ordering. However, there has been less investigation on the selection of the reference frames using the statistic of the block matching amongst reference frames of the multi-view video codecs.

In this paper the contribution of different reference frames in block prediction (taking into account the spatial and temporal location of the reference frames) is investigated. Based on the statistics of the block matching two sets of reference frame architectures for best coding performance of the codec are proposed. The performance of the H.264/AVC based MVV-codec using the proposed reference frame architectures against the codec when it uses the typical reference frame architectures at low bitrates are assessed. Results show that the application of the proposed reference frames architectures significantly improves the coding performance of the codec. The rest of the paper is organized as follows: in Section II multi-view datasets are introduced. Statistical analysis of block matching amongst reference frames is detailed in Section III. H.264/AVC based multi-view

video codec using the statistical analysis of block matching is presented in Section IV and finally paper will be concluded in Section V.

II. DATASET DESCRIPTION

Four multi-view video datasets, called: Break-dancers, Ballet, Race1 and Exit have been used in this investigation. The description of each dataset is provided in Table I. These multi-view videos (MVV) have different characteristics of motion, and scene complexity [11]. Since this investigation targets MVV transmission for mobile devices, CIF and QVGA size multi-view video datasets were generated from Microsoft, KDDI and MERL multi-view sequences. To achieve this, all the frames of the Microsoft datasets have been filtered using a 5×5 FIR Kaiser low-pass filter (the coefficients of this filter are tabulated in Table II); filtered frames are then down-sampled by a factor of 2 both horizontally and vertically to mitigate the aliasing artifacts; to preserve the external camera parameters the resulting frames were cropped from point $(P_x, P_y)=(120,47)$ for Break-dancers and $(P_x, P_y)=(80,47)$ for Ballet sequences and corresponding CIF size sequences were generated. The resulting RGB frames are finally converted to YUV in full color sampling format 4:4:4. The luminance components of the KDDI and MERL datasets were also filtered and down-sampled generating full color sampling QVGA sizes.

Seven views of the Multi-View Videos (MVVs) are considered for this investigation (view zero to view six). Frames of different views are interleaved using time first ordering generating a single sequence [12].

III. STATISTICAL ANALYSIS OF BLOCK MATCHING AMONGST REFERENCE FRAMES

Statistical analysis for macroblock prediction has been conducted using H.264/AVC based multi-view video codec. All inter-picture coding modes, which include 16×16, 16×8, 8×16, 8×8, 8×4, 4×8 and 4×4 block sizes, in addition to intra-prediction have been enabled for this analysis. Bitrate control is also enabled to reveal the block matching statistics of the multi-view video codec amongst temporal and neighboring reference frames at low bitrate transmission (64 kbps). Two comprehensive statistical analysis measurements of block matching amongst reference frames have been carried out using IPPP configuration. Fig. 1 demonstrates the 21 reference frames, where there are three temporal reference frames named: T_0-T_2, three spatial reference frames at time slice zero named: S_0-S_2 and fifteen spatiotemporal reference frames at time slice T_{i-1} to T_{i-3} named: M_0-M_{14}. In the first set of investigation, the reference frames were sorted in descending order, as shown in Fig. 2-a, the primary results for the statistical analysis for central view of Break-dancers is tabulated in Table III. From Table III, it can be seen that the neighboring reference frames have significant contribution for block matching. Rate-distortion optimization of H.264/AVC

TABLE I. DATASETS DESCRIPTION

Dataset name	Provider	Frame size/ Frame format	Inter-cameras' distance/ Frame rate	Camera setup
Break-dancers	Microsoft	1024×768 4:4:4	20 cm 15 fps	1D/arc
Ballet	Microsoft	1024×768 4:4:4	20 cm 15 fps	1D/arc
Race1	KDDI	640×480 4:2:0	20 cm 30 fps	1D/parallel
Exit	MERL	640×480 4:2:0	19.5 cm 25 fps	1D/parallel

TABLE II. KAISER FIR FILTER COEFFICIENTS

0	0	0.0393	0	0
0	0.0653	0.1077	0.0653	0
0.0393	0.1077	0.1511	0.1077	0.0393
0	0.0653	0.1077	0.0653	0
0	0	0.0393	0	0

enforces the minimum cost for the best reference frame selection, which is a compromise between the actual bitrate and residual error due to the inter-prediction from various reference frames. According to Lagrangian method, the cost function, $J(ref \mid \lambda_{motion})$ is defined by "equation (1)" [13]:

$$J(ref \mid \lambda_{motion}) = SAD(s, r) + \lambda_{motion} \times R(MV, REF) \quad (1)$$

where Sum of Absolute Difference (*SAD*) frame is the prediction error between the current *s*, and, corresponding reference block *r*, λ_{motion} is Lagrange multiplier, *R* is the number of bits required to code both; motion vector (*MV*) and reference frame (*REF*).

In H.264/AVC, the index of all used reference frames are stored in buffer; List 0 and the overhead cost of presenting the reference frame for each macroblock is related to the index position of the reference frame in buffer; List 0, where fewer number of bits are used to address the closer reference frames. From Table III, it is obvious that the distribution of the block matching amongst reference frames is inconsistent with the position of the reference frames in the buffer List 0. Hence, current bit allocation system for representing each macroblock reference frame could cost additional bits, which reduces coding efficiency of the codec. Therefore, ordering the index of reference frames according to their contributions in block-matching could improve the coding performance of the codec. In the second set of experiments, the reference frames were first indexed according to their contributions in block matching using the resulting statistics from the first set of experiments beside their spatial position to the current frame, as shown in Fig. 2-b. Another comprehensive statistical analysis of block matching was performed using the proposed reference frame indexing. The results for Break-dancers sequences are tabulated in Table IV. From Table IV, it can be seen that the temporal frames have higher contribution in block prediction

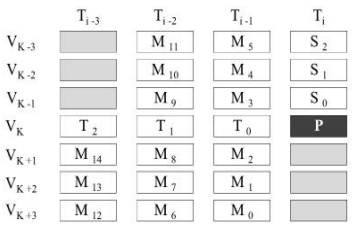


Figure 1. Block diagram of reference frames used in the statistical analysis.

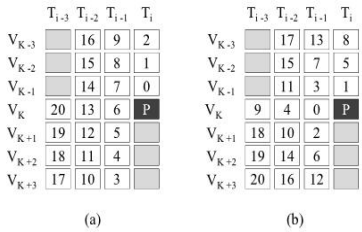


Figure 2. Reference frame indexing: a) descending order and, b) according to their contributions in block matching.

than the spatial frames and also the first temporal frame, T_0, contributes more than the adjacent spatial reference frame, S_0, in block prediction. This finding matches the fact that temporal correlations are higher than the spatial correlations. It reveals that coding performance of the MVV codec could be increased by using the proposed reference frame indexing.

The statistic of Skip and Intra-Prediction for using the first set of indexing reference frames are 40% and 7.6%, while for using the second indexing reference frames are 55.8% and 5.82%, respectively. It can be realized that the percentage of the macroblocks using the Skip mode prediction has been increased by 15.8%. Since, encoded skipped macroblock cost a single bit to signal this mode instead of sending its prediction information. Additional coding performance may be achieved by using the proposed reference frame indexing. It can also be seen that the percentage of the Intra-coded macroblock has been reduced by 1.78%. Since Intra-Prediction macroblocks encoding is more costly than other encoding modes. It implies that the application of the proposed reference frame indexing could improve the achievable coding performance of the multi-view codec.

IV. H.264/AVC BASED MULTI-VIEW VIDEO CODEC USING THE STATISTICS OF BLOCK MATCHING

The outcomes of the statistical analysis for the proposed reference frame indexing have been considered in the selection of the reference frames and also in indexing the reference frames in buffer List 0. Fig. 3-a and Fig. 3-b, show two proposed reference frame architectures for coding multi-view videos using six and four reference frames. In Fig. 3 numbers in each block represents its frame reference indexing and V_k,

TABLE III. STATISTICS OF BLOCK MATCHING AMONGST REFERENCE FRAMES USING THE DESCENDING ORDER FRAME INDEXING.

	T_{i-3}	T_{i-2}	T_{i-1}	T_i
View 0	n/a	0.02	0.087	0.87
View 1	n/a	0.03	0.22	1.95
View 2	n/a	0.37	0.95	66
View 3	0.67	1.8	16.85	P
View 4	0.2	0.57	5.06	
View 5	0.08	0.26	1.67	
View 6	0.09	0.26	2.03	

TABLE IV. STATISTICS OF BLOCK MATCHING AMONGST REFERENCE FRAMES USING THE PROPOSED FRAME INDEXING.

	T_{i-3}	T_{i-2}	T_{i-1}	T_i
View 0	n/a	0.02	0.07	0.29
View 1	n/a	0.01	0.1	1.36
View 2	n/a	0.15	1.18	23.61
View 3	0.5	1.38	67.64	P
View 4	0.1	0.22	1.93	
View 5	0.08	0.17	0.73	
View 6	0.04	0.1	0.32	

$V_{k\pm1}$ and V_{k-2} represent the current and its three corresponding neighboring views. In order to evaluate the performance of using the proposed reference frame architectures, four MVVs sequences were taken (described in section II). The H.264/AVC based multi-view video codec using the proposed reference frame architectures and three reference architectures were applied to these multi-view sequences. The first (Typical-A) and second (Typical-B) prediction structures are typical structures for IPPP (based on the outcome on [6, 7]) where the first gives higher priority to temporal reference frames while the latter places the spatial reference frame first in List 0. The third prediction structure represents Fecker's prediction structure as in [9].

The Peak Signal to Noise Ratio (*PSNR*) measurement was used to assess the quality of the reconstructed luminance components of decoded sequences. Fig. 4-a to 4-d, show the resulting *PSNR* when coding Ballet, Break-dancers, Race1 and, Exits MVVs at 32, 64 and 96 kbps. From Fig. 4, it can be seen that the application of the proposed reference frame architectures and indexing improves the coding performance significantly of the codec compared to Fecker's prediction structure (up to 2.3 dB) using six reference frames while it improves coding gain using four reference frames by up to 0.43 dB and 0.83 dB compared to Typical-A and Typical-B configurations respectively. It can be observed that the proposed reference frame architecture with 4 reference frames gives superior performance to other architectures at 32 KBPS while the proposed reference frame architecture with 6 reference frames delivers superior coding performance at higher bitrates (96 KBPS).

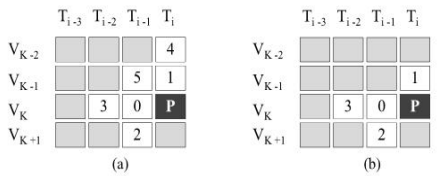


Figure 3. Proposed Reference frame architectures using: a) 6 reference frames and b) 4 reference frames.

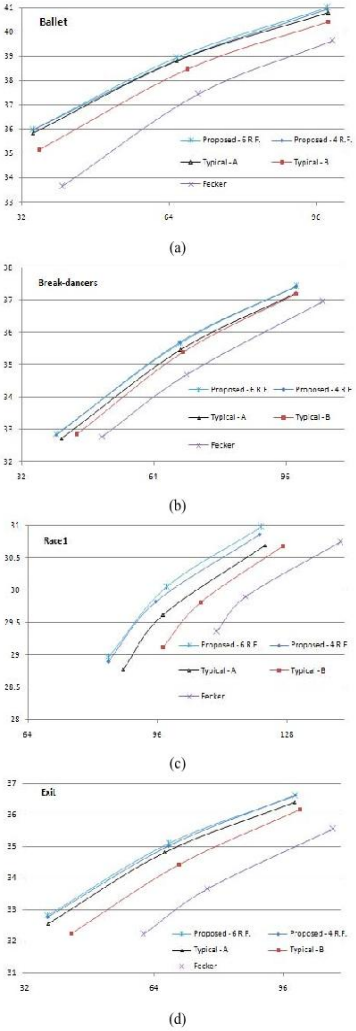


Figure 4. Coding performance of the codec using the proposed reference frame architecture using 6 and 4 reference frames and three different reference frame architectures for: a) Ballet, b) Break-dancers, c) Race1 and d) Exit.

V. CONCLUSIONS

H.264/AVC based multi-view video codec with two sets of reference frame architectures was found to give superior coding performance compared to the same codec using two widely used reference frame architectures. Two sets of statistical analysis of block matching amongst reference frames were carried out by taking into consideration the spatial and temporal location of the reference frames. Results showed that there is a strong link between the numbers of reference frames, target bitrates and coding performance of the codec. Results revealed that at very low bitrates application of reference frame architecture with smaller number of frames is preferred while at higher bitrates the application of the architecture with more reference frames generates superior performance.

REFERENCES

[1] A. Smolic, "3D video and free viewpoint video - From capture to display," Pattern Recognition, Elsevier, vol. 44, no. 9, pp. 1958–1968, Sep. 2011.

[2] S. Yuehou, Y. Mei and P. Zongju, "A Fast Multi-reference Frame Selection Algorithm for Multiview Video Coding," Journal of Multimedia, Academy Publisher, vol.5, no.4, pp.369-376, Aug. 2010.

[3] M. Abdoli, M. Soryani, and a. F. A. Modarres, "The Impact of View Spacing in Multi-view Video Compression Efficiency," in 7th International Conference on Information Technology: New Generations, pp. 1314–1315, Apr. 2010.

[4] S. A. Fezza, K. M. Faraoun, and S. Ouddane, "A comparison of prediction structures for multi-view video coding based on the H.264/AVC standard," in International Workshop on Systems, Signal Processing and their Applications, WOSSPA, pp. 111–114, May 2011.

[5] 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, Apr. 2011.

[6] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient Prediction Structures for Multiview Video Coding," IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 11, pp. 1461–1473, Nov. 2007.

[7] A. Kaup and U. Fecker, "Analysis of multi-reference block matching for Multi-View Video Coding," Proceedings of 7th Workshop Digital Broadcasting, Germany, pp. 33 -39, Sept. 2006.

[8] C. Bilen, A. Aksay, and G. Akar, "A Multi-View Video Codec Based on H.264," in 2006 International Conference on Image Processing, pp. 541–544, Oct. 2006.

[9] U. Fecker and A. kaup, "H.264/AVC-Compatible Coding of Dynamic of Light Fields using Transposed Picture Ordering," EUSIPCO 2005, Turkey, pp. 1-4, Sep. 2005.

[10] H. Said, A. Sheikh Akbari, and A. Malik, "H.264/AVC based stereoscopic video coding scheme using the statistical analysis of the block matching," 36th International Conference on Telecommunications and Signal Processing (TSP), July 2013, in press.

[11] Y. Zhang, S. Kwong, G. Jiang, and H. Wang, "Efficient Multi-Reference Frame Selection Algorithm for Hierarchical B Pictures in Multiview Video Coding," IEEE Transactions on Broadcasting, vol. 57, no. 1, pp. 15–23, Mar. 2011.

[12] Y. Chen, Y.-K. Wang, K. Ugur, M. M. Hannuksela, J. Lainema, and M. Gabbouj, "The Emerging MVC Standard for 3D Video Services," EURASIP Journal on Advances in Signal Processing, no. 1, pp. 1-13, 2009.

[13] S.-H. Jung, W.-J. Park, and T.-Y. Kim, "Fast Reference Frame Selection with Adaptive Motion Search Using RD Cost," in 2012 Spring Congress on Engineering and Technology (S-CET), pp. 1–4, May 2012.

EUSIPCO 2013 1569743887

AN ADAPTIVE REFERENCE FRAME RE-ORDERING ALGORITHM FOR H.264/AVC BASED MULTI-VIEW VIDEO CODEC

[1]Hany Said, [2]Akbar Sheikh Akbari and [3]Mansour Moniri

Faculty of Computing, Engineering and Sciences
Staffordshire University, Beaconside, Stafford, UK
Email: {[1]h.h.said, [2]a.s.akbari, [3]m.moniri}@staffs.ac.uk

ABSTRACT

This paper proposes an adaptive reference frame re-ordering for H.264/AVC based multi-view video codecs. The algorithm relies on statistical analysis of block matching among reference frames at low bitrate. The coded macroblocks are statistically analysed and the corresponding order for reference frames is then determined. The adaptive reference frame re-ordering algorithm is evaluated for two scenarios. In the first scenario, the multi-view videos are coded using a prediction structure with a number of reference frames. In the second scenario, a video sequence that contains several scene changes is coded. The proposed algorithm has been tested using two different prediction structures for both scenarios. The measurements were carried out on four standard multi-view datasets in addition to a sequence that contains several scenes changes. Results show that the application of the proposed reference frame re-ordering algorithm significantly saves up to 6.2% of the bitrate when coding a sequence with multiple scene changes and up to 0.2 dB when coding a sequence using multiple reference frames at low bitrate.

Index Terms— H.264/AVC, Multi-view video codec, statistical analysis, reference frames re-ordering, scene change

1. INTRODUCTION

Multi-view videos (MVVs) enable the viewer to watch these type of videos from different view-points as in free viewpoint TV (FTV) or enjoys perceiving scene depth through watching 3D videos as in three-dimensional TVs (3D-TVs) [1]. These MVVs are generated by capturing the same scene using multiple synchronized cameras at different positions and view-points [2]. Multi-view videos (MVV) contain several videos; their sizes are proportional with the number of views and resulting in huge amount of visual data which need to be compressed efficiently to enable the applications of FTV and 3D-TV.

Since the cameras filming the same scene, multi-view videos share significant amount of correlations among their views [3]. These correlations enable H.264/AVC to code MVVs efficiently through extending its coding property of multiple reference frames to exploit efficiently these correlations [4-10].

It can be seen from the literature that the H.264/AVC based MVV Codecs (MVCs) use different prediction architectures with different number of reference frames and reference frame orderings to improve their coding efficiency. Reference frame selection entails coding the current frame using previous decoded frames. These decoded frames are frames that belong to the current view (temporal reference frame) or neighbouring views (spatial and spatiotemporal reference frame) [4-10]. Reference frames (RFs) ordering reflect the way that the reference frames is placed inside H.264/AVC Decoded Picture Buffer (DPB) where few numbers of bits are used to address the closer reference frames inside this buffer (Buffer list0 is used when coding P-frames and, buffers list0 and list1 for compressing B-frames). A number of H.264/AVC based MVCs with different static RFs ordering for coding P-frames have been reported in the literature [6-10]. Temporal RFs are placed either at the beginning of list0 (e.g. [8] and [10] are depicted in Figure 1-a and, 1-b respectively), or at the end of the buffer (as shown in Figure 1-c [10]). *Fecker and Kaup* ordered RFs in opposite direction of the coding order [5] while temporal, spatial and spatiotemporal RFs are placed in an interleaved manner inside buffer as in [6, 7]. Dynamic RFs ordering for stereoscopic video coding was proposed by *Hong and Yu* [9]. Their algorithm re-orders the RFs when the number of skipped macroblocks increases. Although this algorithm efficiently encodes the stereoscopic videos, it may not meet the requirements of the real-time applications as each frame is encoded twice.

In this paper, an adaptive RFs re-ordering algorithm for multi-view video coding is proposed that encode each frame once. The proposed algorithm determines the significance of each reference frame in terms of how much it has been used as a reference in predicting blocks. Hence an analysis of block matching among the reference frames is performed to reveal the statistics of block matching. Based on the statistics of block matching for each frame, reference frames are adaptively re-ordered such that the significant references

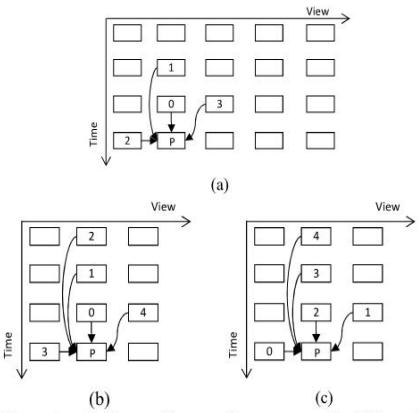


Figure 1-a Reference frame orders proposed by Bilen *et al.* (mode 3) [8], Figures 1-b and 1-c are modes 3 and mode 1 respectively that represent reference frame orders proposed by Sheikh Akbari *et al.* [10].

frames are placed first in DPB. Performance of the MVC using the propose RFs re-ordering is evaluated against the use of a statistic RFs order in two different scenarios. The first scenario is concerned with coding standard MVVs [11] and the second is concerned with coding a sequence with multiple scene changes. The performance of the H.264/AVC based multi-view video codec using the proposed algorithm is applied on prediction structure proposed in [10] for the first scenario and on prediction structure proposed in [8] for the second scenario. Results indicate the merit of the proposed RFs re-ordering in dealing with scene changes. The rest of the paper is organized as follows: in Section 2 MVV datasets are introduced. Section 3 briefly justifies the necessity of using fix RFs ordering or adaptive RFs re-ordering. Adaptive reference frame re-ordering algorithm is presented in Section 4. Experimental results are given in Section 5 and finally paper is concluded in Section 6.

2. DATASET DESCRIPTION

Four multi-view video datasets have been used in this investigation. The description of each dataset is provided in Table I. These MVVs are captured using eight cameras and they have different characteristics of motion, disparity and scene complexity [12]. Since this investigation targets coding multi-view video at low bitrate transmission, MVV datasets of CIF and QVGA size were generated from Microsoft, KDDI and MERL multi-view sequences. To achieve this, all the frames of the Microsoft datasets have been filtered using a 5×5 FIR Kaiser low-pass filter (the coefficients of this filter are tabulated in Table II); filtered frames are then down-sampled by a factor of 2 both

Table I Datasets Description

Dataset Name	Provider	Frame size / Frame format	Camera setup	Inter-cameras' distance
Break-dancers	Microsoft	1024×768 4:4:4	1D/arc	20 cm
Ballet	Microsoft	1024×768 4:4:4	1D/arc	20 cm
Race1	KDDI	640×480 4:2:0	1D/ parallel	20 cm
Exit	MERL	640×480 4:2:0	1D/ parallel	19.5 cm

Table II KAISER FIR FILTER COEFFICIENTS

0	0	0.0393	0	0
0	0.0653	0.1077	0.0653	0
0.0393	0.1077	0.1511	0.1077	0.0393
0	0.0653	0.1077	0.0653	0
0	0	0.0393	0	0

horizontally and vertically then the resulting frames are cropped from point (P_x, P_y)=(120,47) for Break-dancers and (P_x, P_y)=(80,47) for Ballet sequences and corresponding CIF size sequences are generated [7]. The resulting RGB frames are finally converted to YUV in full colour sampling format 4:4:4. The luminance components of the KDDI and MERL datasets are also filtered and down-sampled generating full colour sampling QVGA sizes. Frames of different views are interleaved using time first ordering to generate a single sequence [12]. A sequence of QVGA size with different multi-view scenes is generated by interleaving the previous MVVs together. Microsoft datasets are further down-sampled in order to match QVGA resolution size. The first six frames from each view within MVVs are used to generate a MVV sequence where sixteen consecutive frames from each video are concatenated to a MVV sequence, thus the resulting sequence contains 192 frames.

3. DO FRAMES USE SAME OR DIFFERENT REFERENCE FRAME ORDERING IN MULTI-VIEW VIDEO CODING?

In a H.264/AVC based multi-view video codecs, the order of RFs is fixed through coding the entire MVVs. This section investigates whether frames should use the same order of RFs or should they follow different RFs orders. A statistical analysis of block matching among reference frames has been conducted using H.264/AVC based MVV codec using a prediction structure depicted in Figure 2. This analysis determines the contribution of each RF for

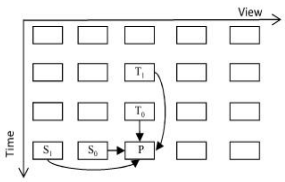


Figure 2 the prediction structure used in investigating reference frame order

Table III Six cases for reference frames orders

Case Label	Ref_0	Ref_1	Ref_2	Ref_3
A	T_0	S_0	S_1	T_1
B	T_0	S_0	T_1	S_1
C	S_0	T_0	S_1	T_1
D	S_0	T_0	T_1	S_1
E	T_0	T_1	S_0	S_1
F	S_0	S_1	T_0	T_1

Table IV shows labels which reflect the appropriate order of reference frames for the coded Break-dancers.

t_i / V_k	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}
V_2	C	C	C	C	D	C	D	E	D	C	C
V_3	B	B	B	B	B	B	A	B	B	B	B
V_4	C	D	C	C	C	C	C	C	C	C	C
V_5	A	A	A	C	C	C	C	C	C	C	C
V_6	C	C	C	C	C	C	C	C	C	C	C

Table V shows labels which reflect the appropriate order of reference frames for the encoded Ballet.

t_i / V_k	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}
V_2	A	C	D	B	B	B	B	B	B	B	A
V_3	B	B	B	A	B	A	A	A	A	A	B
V_4	A	B	B	B	A	A	B	A	A	B	A
V_5	B	A	B	A	D	B	D	D	C	B	C
V_6	B	A	B	C	C	C	C	C	C	C	C

predicting P-frame using all block sizes. All inter-picture coding modes and intra-prediction have been enabled. Bitrate control is enabled to encode the given MVV at low bitrate (64 Kbps).

The basic idea beyond this section is to reveal the order of RFs after encoding the P-frame using this order; T_0, T_1, S_0 and S_1. The statistic of the block matching amongst RFs is calculated and used to sort the RFs in descending order.

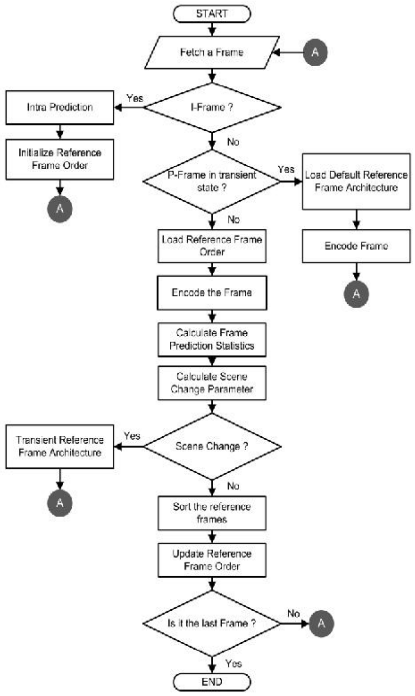


Figure 3 Adaptive Reference Frames Re-ordering Algorithm

The sorted RFs are then given a label. These labels are tabulated in Table III, where there are six different RFs orders starting from Label A to Label F and Ref_i represent a temporal T- or Spatial S-reference frame.

This investigation has been applied on Break-dancers and Ballet using the first seven views. The first two views; V_0 and V_1; are not involved in this analysis due to unavailability of some reference frames (e.g. S_0 and S_1). Tables IV and V, show the suitable reference frames order in terms of "labels", based on the statistics of block matching among four reference frames for the first 55 frames from time step t_2 to t_{12}. It is worth to mention that RFs order labelled by 'A' and 'B' are similar because their first two RFs are the same (T_0 then S_0) and they always have the most contribution of block matching prediction (the same concept applies to labels 'C' and 'D'). The shaded cells in Table IV and V show consecutive frames within the same view (temporal frames) that should be coded using different RFs orders. Also, it can be inferred that the suitable RFs order would be predicted in most cases, using previous frames within the same view.

4. ADAPTIVE REFERENCE FRAME RE-ORDERING ALGORITHM

Section 3 shows that the order of RFs is predicted using the corresponding information from the recent temporal frames. The flow of the proposed algorithm is presented in Figure 3. For a P-frame, it checks first if the frame is located in a position where partial of RFs are available (transient state e.g. all P-frames in first time slice; t_0). In this stage, the algorithm uses predefined prediction architecture to encode the frame where the prediction structure involves available RFs with their initial order. In a non-transient scenario, the algorithm loads the corresponding order of RFs then encodes the P-frame using that order. After that, the algorithm loops on all its macroblocks to compute the block matching statistics among all RFs. When there is no scene change, the algorithm orders the reference frames based on their block matching statistics and its new order will be stored and applied to the next temporal frame.

When scene changing, the majority of frame's macroblocks in the new scene are intra-predicted. Hence the algorithm is relying on the number of intra-coded macroblocks to detect scene changes. If the percentage of intra-predicted macroblocks exceeds certain threshold (60%) [13], then the following P-frames will use similar RFs order to the corresponding P-frames in transient state (e.g. following frame in coding order will use RFs order where spatial RFs are placed first in DPB). In other word, following frames that are located within the same time slice when scene changes, will use RFs order where spatial reference frames are placed first in list0.

5. EXPERIMENTAL RESULTS

The proposed algorithm has been evaluated in encoding MVVs, (Break-dancers, Ballet, Race1 and Exit) and also encoding a sequence that contains a number of scene changes at low bitrate. The proposed algorithm has been implemented using prediction structures reported by Sheikh Akbari *et al*. [10] and Bilen *et al*. [8] as they clearly highlighted the order of the selected reference frames in their reported prediction structures.

In the first scenario, the algorithm uses the prediction structures proposed in [10] for coding four different MVVs at low bitrates. This prediction structure contains five reference frames with two different reference frame orders. Figure 3-c presents the first reference frame order where spatial and spatiotemporal RFs have higher priority than the temporal frames (Mode 1 in [10]). Figure 3-b places temporal reference frames in the beginning of the other reference frames (Mode 3 in [10]). The proposed algorithm starts with the same order of reference frames that was suggested in each Mode. P-frame located in time slice below t_3 will be coded using the available reference frames (transient state). After t_3, the algorithm starts to adapt the reference frames re-ordering dynamically. Figure 4 shows

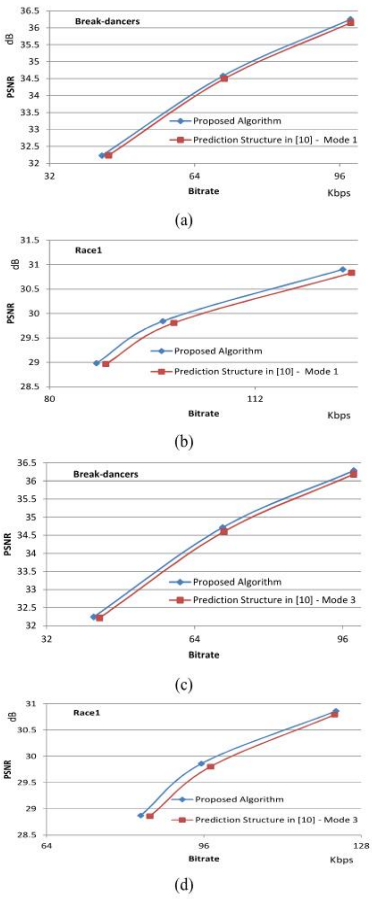


Figure 4. Coding performance of the MVC using the proposed algorithm on the prediction architectures proposed by Sheikh Akbari *et al*. [10] using: a-b) Mode 1 and c-d) Mode 3.

the coding performance of the MVC using the proposed adaptive re-ordering algorithm in coding Break-dancers and Race1 MVV datasets in comparison to RFs order proposed in [10]. From Figure 4, it can be seen that the proposed algorithm gives higher coding performance compared to the use of static RFs orders (up to 0.2 dB).

In the second scenario, the proposed RFs re-ordering algorithm and the prediction structure reported in [8] (Mode 3 is shown in Figure 1-a) are used to code a sequence with scenes changes. Results are shown in Figure 5. From figures 5 and 6, it can be seen that the proposed algorithm gives

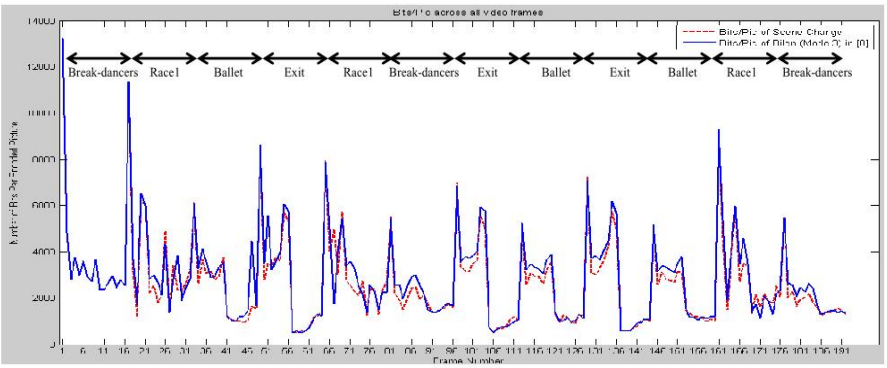


Figure 5. Number of bits per coded picture when using prediction structure proposed in [8] and the proposed algorithm.

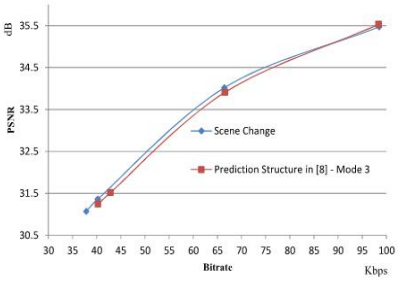


Figure 6. Coding performance for the proposed algorithm using the prediction structure proposed by Bilen *et al.* [8].

slightly higher coding performance compared to the use of static RFs order (as shown in Figure 6), at the same time it saves significant bitrates, up to 6.2%.

6. CONCLUSIONS

An adaptive reference frames re-ordering was proposed. The proposed algorithm updates the reference frame orders adaptively using the statistics of block matching. The proposed re-ordering algorithm gives superior coding performance compared to the state of arts (up to 0.2 dB). In addition, it efficiently re-orders reference frames when dealing with scene changes and saves bitrates of up to 6.2%.

7. REFERENCES

[1] M. Tanimoto, "FTV: Free-viewpoint Television," *Signal Processing: Image Communication*, vol. 27, no. 6, pp. 555–570, Jul. 2012.

[2] A. Smolic, "3D video and free viewpoint video - From capture to display," *Pattern Recognition*, Elsevier, vol. 44, no. 9, pp. 1958–1968, Sep. 2011.

[3] J. Seo and K. Sohn, "Early disparity estimation skipping for multi-view video coding," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, Feb. 2012.

[4] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the Stereo and Multi-view Video Coding Extensions of the H.264/MPEG-4 AVC Standard," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626–642, Apr. 2011.

[5] U. Fecker and A. Kaup, "H.264/AVC-Compatible Coding of Dynamic Light Fields Using Transposed Picture Ordering," *EUSIPCO 2005*, Antalya, Turkey, Sept. 2005.

[6] H. Said, A. Sheikh Akbari and, M. M. Abbas Malik, "H.264/AVC based Stereoscopic Video Coding Scheme using the Statistics of Block Matching," accepted for publication in 36th International Conference on Telecommunications and Signal Processing TSP 2013, July 2013.

[7] H. Said and, A. Sheikh Akbari, "H.264/AVC Based Multi-view Video Codec using the Statistics of Block Matching," accepted for publication in 55th International Symposium ELMAR-2013, Sept. 2013.

[8] C. Bilen, A. Aksay, and G.B. Akar, "A Multi-View Video Codec Based on H.264," *in 2006 International Conference on Image Processing*, pp. 541–544, Oct. 2006.

[9] S. Hong and Y. Yu, "Dynamic reference frame reordering for frame sequential stereoscopic video encoding," Patent, US 20110109721, Sony Corporation, Jul 2012.

[10] A. Sheikh Akbari, N. Canagarajah, D. Redmill, D. Agrafiotis, "A Novel H.264/AVC Based Multi-View Video Coding Scheme," *3DTV Conference 2007*, pp.1-4, 7-9 May 2007.

[11] Y. Zhang, S. Kwong, G. Jiang, and H. Wang, "Efficient Multi-Reference Frame Selection Algorithm for Hierarchical B Pictures in Multi-view Video Coding," *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 15–23, Mar. 2011.

[12] Y. Chen, Y.-K. Wang, K. Ugur, M. M. Hannuksela, J. Lainema, and M. Gabbouj, "The Emerging MVC Standard for 3D Video Services," *EURASIP Journal on Advances in Signal Processing,* vol. 2009, no. 1, 2009.

[13] J. Brandt, J. Trotzky, L. Wolf, "Fast Frame-Based Scene Change Detection in the Compressed Domain for MPEG-4 Video," *in Proc. of Next Generation Mobile Applications, Services and Technologies, 2008. NGMAST '08*, pp.514-520, Sept. 2008.