Reliable Streaming of Stereoscopic Video Considering Region of Interest

by

Ehsan Rahimi, **M.Sc.**

A thesis proposal submitted to the
Faculty of Graduate and Postdoctoral Affairs
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario
December, 21

Abstract

3D video applications are growing increasingly common as the required infrastructure and technology to stream 3D video becomes more predominant. However, the quality of displayed videos may fluctuate due to packet failure as an integral part of either wired or wireless streaming networks. Therefore, more robust methods of video streaming have always been fascinating to show more favourable efficiency outcomes.

This thesis first examines different video streaming techniques and compares the pros and cons of each technique. It then introduces a new streaming method that applies to 3D video for live video streaming applications especially for a sporting event or other live video applications.

To this end, the thesis describes how a 3D video is captured and represented, and how humans perceive the 3D scene. Considering the pros and cons of current video streaming techniques and intended applications, the proposed method introduces a new multiple description coding (MDC) method focusing on interesting objects of the scene, called the region of interest (ROI). It is worth mentioning that a new technique, using the scene's depth information, is used to extract the ROI. This technique is not as complex as learning algorithms are, and there is no need to train the algorithm. Since the human eye is more sensitive to objects than pixels, this method can also provide better performance from the point of subjective assessment (which is out of focus of this thesis) because the proposed method focuses on important objects of the scene and assigns more bandwidth to them.

To
my beloved parents, Zahra and Mohammadali,
my lovely wife, Elham,
and my sweet little daughter, Elsa.

Acknowledgments

First and foremost, I would like to express my sincere appreciation to my supervisor, Professor Chris Joslin, for his unrelenting support, invaluable guidance, and constructive suggestions throughout my studies. It was a great privilege and honour to work and study under his guidance.

Heartfelt gratitude and love go to my parents, Zahra and Mohammadali, for providing me with abundant support and continuous encouragement throughout my years of study.

Last, but not least, I wish to express my infinite appreciation and love to my wife, Elham. Without her endless patience and unfailing support, completing this thesis would not have been possible.

Table of Contents

List of Tables

List of Figures

Acronyms

Acronym	Meaning
2D	Two Dimensional
3D	Three Dimensional
3DTV	Three Dimensional TeleVision
5G	Fifth Generation
ARQ	Automatic Repeat reQuest
AVC	Advanced Video Coding
BL	Base Layer
CDF	Cumulative Density Function
CGS	Coarse Grain Scalability
CIF	Common Intermediate Format
CV	Coefficient of Variation
DASH	Dynamic Adaptive Streaming over HTTP
DCT	Discrete Cosine Transformation
DPCM	Differential Pulse Code Modulation
DVD	Digital Versatile Disc
EL	Enhancement Layer
ERC	Error Resilient Coding
FEC	Forward Error Coding

FGS	Fine Grain Scalability
FMC	Flexible Macrobloc Ordering
FPS	Frame per second
FTV	Free viewpoint TeleVision
GOP	Group of Picture
HD	High Definition
HEVC	High Efficiency Video Coding
HBDA	Hierarchical Block Division Algorithm
HTTP	Hyper Text Transfer Protocol
IP	Internet Protocol
ITU	International Telecommunication Union
LC	Layered Coding
LOT	Lapped Orthogonal Transformation
LTE	Long Term Evolution
MDC	Multiple Description Coding
MDSQ	Multiple Description Scalar Quantizer
MDTC	Multi Description Transform Coding
MGC	Medium Grain Scalability
MPEG	Moving Picture Experts Group
MS-MDC	Multistate Stereo-Multiple Description Coding
OFDM	Orthogonal Frequency Division Multiplexing
PAL	Phase Alternating Line
PCT	Pairwise Correlating Transformation
PDF	Probability Density Function

PSS	Polyphase SubSampling
PSNR	Peak Signal-to-Noise Ratio
PV	Pixels Variance
QCIF	Quarter Common Intermediate Format
ROI	Region Of Interest
RTT	Round Trip Time
MC	Motion Compensation
RVLC	Reversible Variable Length Coding
SHVC	Scalable High efficiency Video Coding
SNR	Signal to Nose Ratio
SSIM	Structural Similarity Index)
SS-MDC	Spatial scaling Stereo-Multiple Description Coding
SVC	Scalable Video Coding
TV	Television
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
UHD	Ultra High Definition
VSP	View Synthesis Prediction
VOD	Video On Demand

Chapter 1

Introduction

1.1 Overview

3D displays have long been favoured by moviegoers. Since September 27, 1922, when the first 3D film, "The Power of Love," was displayed at the Ambassador Hotel theater in Los Angeles, viewers have been drawn to the enhanced experience of 3D displays [1]. The film producer in that early case was Harry K. Fairall, one of the pioneers of stereoscopy. Although the film is no longer available, it was a projected dual-strip in the now-classic red and green anaglyph format. Until the last decade, 3D movies were limited to public displays because of hardware and software limitations. New technological advancement in hardware and software over the past few decades has made it possible for the ordinary family to watch 3D movies at home.

According to [2], 3D videos can be represented in three different ways: panoramic video, stereoscopic video, and multiview video. Briefly, panoramic video or omnidirectional video is "an extension of a 2D video display plane into a spherical surface in order to make audience able to see around the object and transfer the feeling of being surrounded in a 3D scene" [2]. Stereoscopic video is a subset of multiview video, since stereoscopic video captures only two adjacent views, compared to multiview video, which captures a multiplicity of views. Stereoscopic video requires less resources than multiview video. The simpler camera adjustments, lower disk space, reduced processing power, and lower bandwidth requirements of stereoscopic video make it the most popular of the three 3D video technologies. In general, stereoscopic video can be produced using one of these methods: 2D to 3D conversion, dual-camera configuration, or a 3D/depth-range camera, which will be explained in Chapter 2 [2, 3].

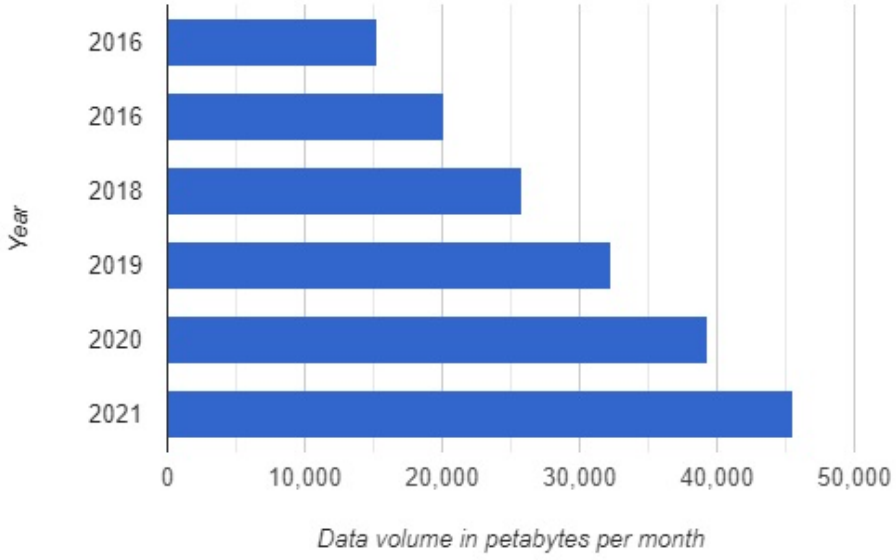


Figure 1.1: Consumer internet video traffic in North America 2016-2021.

1.1.1 General Video Streaming

The development of digital communication systems has accelerated the growth of wired and wireless communications to the extent that everyone is now able to use multimedia applications, such as TV on-demand, video conferencing, online games, social media applications, etc. on mobile devices everywhere. Due to such ubiquitous accessibility of multimedia communication, bandwidth demands have dramatically increased and therefore, old communication technologies are no longer able to support transmission. As can be seen in Figure 1.1, it is estimated that the video content data traffic in North America will triple by the end of next year compared to five years ago (from 15K petabytes per month in 2016 to 45K petabytes per month in 2021) [4].

As of the first quarter of 2020, Netflix had over 182 million paying subscribers which is about 60% of total subscribers streaming services. According to The Wall Street Journal [5], Netflix statistics reveal that the company's subscribers increased by over 16 million subscribers across the world. Such an increase in the number of subscribers made Netflix the most-used video streaming service of 2020. One reason

for such an increase could be the 2020 coronavirus pandemic, which highlights the importance of video streaming in the current lock-down situation. Deloitte also claims that online media streaming benefits from the coronavirus pandemic [6].

To add even more pressure to bandwidth capacity, ultra HD (UHD) TV and 3D/multiview videos are becoming more popular among multimedia users. Cisco estimates the number of installed UHD TV sets will have doubled by 2023 compared to those installed in 2018 [7].

Furthermore, the live streaming industry is growing very fast. For example, a report shows that "live video grew by 93%, with an average viewing time of 26.4 minutes per session" [8]. In addition, quality has always been important for users of online streaming services. According to [9] platforms with lower quality videos run the risk of losing about 25% of their revenue. This emphasizes the importance of correcting errors that happen in communication channels when a video is streamed.

1.1.2 Sport Event Streaming

One popular video broadcast service is sport event streaming, which has attracted large audiences.

According to [10], most TV viewers in the United States watch live sport competitions. As shown in Figure 1.2, about 154 million observers in the United States watched live sport events at least once per month in 2019. This number is expected to rise to over 160 million by 2024.

As another example, it can be claimed that the FIFA World Cup, played every four years, is the most widely watched sporting event in the world. For the most recent FIFA World Cup event, the 2018 FIFA World Cup tournament, stats show that more than 3.5 billion people tuned into the competition [11]about half of the world population in 2018 [12]. Furthermore, the total revenue was projected to be six billion dollars, an increase of 25 percent over 2014 (about $5 billion total revenue). According to FIFA, the broadcast revenue of the 2018 World Cup was projected to rise to three billion dollars [11]. This tournament is only one example of many sporting events that take place every year around the world to emphasize how sport event streaming is beneficial and worthy of investment by video streaming companies.

In addition to the large revenue and huge number of viewers, the largest share of fans are young people, who are no longer passive and limited to broadcast TV. Today, young fans are looking for immersive experiences that let them interact more. They

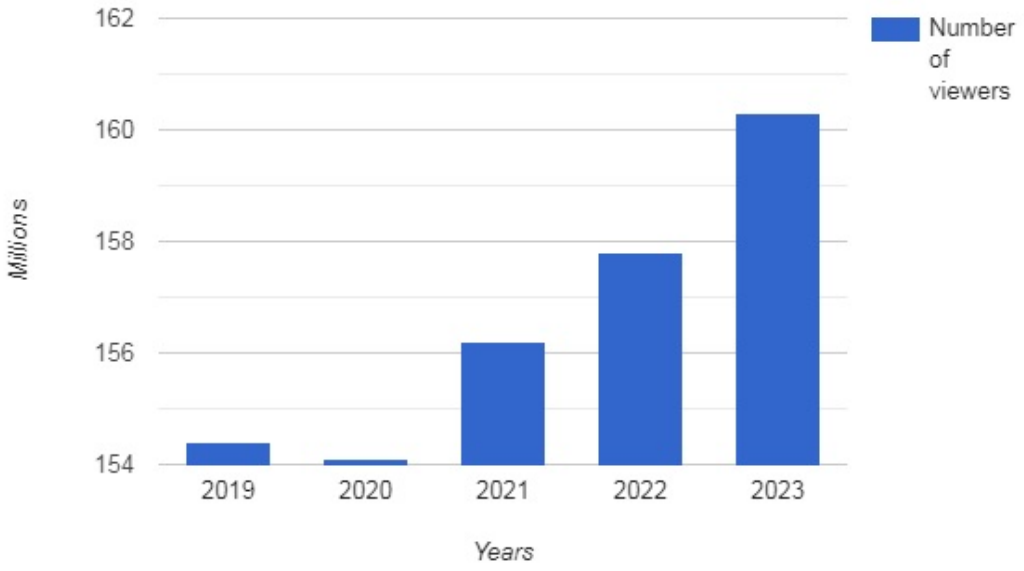


Figure 1.2: Live sport event streaming in the United States in 2020.

are used to working with interactive media and virtual reality, which gives users the feeling of being part of the event. Virtual reality can improve the user experience of a sporting event. Some systems allow the users to walk through a stadium, which helps them when purchasing a ticket to actually attend the event. Companies like Facebook, Google, Microsoft, and YouTube offer virtual reality services, which are gaining increasing attention; however, it is still quite challenging and expensive to generate high-quality virtual reality content [13].

The coronavirus pandemic has also increased the attention given to sport event streaming, as some fans would rather not attend a stadium physically. Pandemic concerns aside, fans can save money on parking tickets, commuting expenses, and accommodation. Even after the current lock-downs lift, it is expected that stadiums will not be able to host competitions with their stands at full capacity. Social distancing will be maintained even after the spread of this virus is contained. Sport streaming also gives the audience the ability to follow synchronous competitions, which are usually common in large tournaments. In such situations, 3D/multiview streaming can

increase the enjoyment and interaction of users when watching sport competitions compared to formerly 2D streaming.

Producing 3D/multi-view video requires a multiple-camera setup, which is quite expensive; however, sport events are currently captured via multiple high-end cameras from different positions. For instance, in the FIFA Womens World Cup 2015, matches were captured and broadcast using more than twenty cameras [13]. Such a set up is another reason that this thesis targets sport event streaming as the major application for its discoveries.

1.2 Motivation

Smolic and Kimata described 3D videos in technical terms as "geometrically calibrated and temporally synchronized (group of) video data or image-based rendering using video input data". [14]; They also added in [14], that a 3D video can be defined as video based rendering. Therefore, 3D videos provide the sensation of depth by adding a depth dimension to former 2D videos at the expense of increasing demand for resources, such as processing power, memory, and bandwidth. In this regard, first, consumers must upgrade most of their current equipment. For example a 3D-capable HDTV is the obvious prerequisite. Historically, the next item had been a 3D capable Blu-ray player; although, with the advent of new video coding standards and also high data rate communications infrastructure, 3D videos are accessible using VOD (Video On Demand) services like VuDu (which is now offering 3D movies in the USA).

The main challenge of these emerging technologies is to adapt them into existing communication infrastructure. Due to limited resources, channel failure has always been seen as an inherent difficulty in multimedia communication and needs to be mitigated. To this end, video coding standard organizations and researchers have recently introduced more efficient 3D video coding standards and methods.

To see how limited resources affect users' experience, we need to explain how 3D videos are captured, encoded, and streamed (see Chapter 2). Briefly, every 3D video frame contains depth information of objects in addition to the colour information that was already available in 2D videos. So, to store or stream the depth information (or multi-view displays), more memory and bandwidth are required.

Even though new memory technology promises fast and high volume production,

it is still challenging to store the enormous volume of 3D video data effectively. Currently, 4K TV boxes, like Apple TV, Roku, Fire TV, or other Android-based TV boxes, with the required storage, RAM, and processing power to support 4K resolution display, are accessible everywhere.

In addition to memory and processing power restriction, bandwidth limitation, network fluctuation, and unreliable communication are bottlenecks of today's multimedia communication and can affect users' experiences dramatically. However the common packet switching network was invented to deal with the problem of transporting bursty data traffic and is characterized as a best effort service with no guarantee of fixed bandwidth allocation.

As a major solution to tackle the high load of data traffic in the network, video compression is crucial to decrease the demand for bandwidth. For the last 30 years, researchers have tried to introduce new methods of compression, scalability, and adaptation in this regard. From the first digital video coding standard in 1984, i.e., ITU-T H.120, to the most recent one in 2013, i.e., High Efficiency Video Coding (HEVC), several video coding standards have been designed or modified to compress or stream video data efficiently and dynamically [15]. Regarding 3D/multiview, the multiview profile was added to the MPEG2 video coding standard in 1998 [16]. The Joint Video Team (JVT) also added the multiview extension to the H264/AVC in 2003 [17–19]. Although with the advent of HEVC, the major step to delivering an HD video to consumers has been achieved, challenges and opportunities for the researchers still exist to enhance the video streaming services efficiency.

Packet loss occurs when a data packet travelling across a computer network is not delivered to its destination. Packet loss is usually caused by network congestion; however, its source may derive from other factors, such as device performance, software issues, and faulty hardware or cabling. Network congestion increases packet delay and finally results in dropping packets. Bandwidth delay, propagation delay, store and forward delay, and querying delay contribute to network latency. Network routers are usually the most important source of latency on the end-to-end path. The more links and router hops data must traverse, the larger the end-to-end latency can be.

ITU-T Recommendation G.114 [20] recommends a one-way delay of less than 400 ms for general network planning. Delays lower than 400ms for highly interactive tasks, such as voice calls, interactive data applications, and video conferencing make for a

Table 1.1: Response time requirements [21].

Application	One way delay	reference
Conversational voice	<150 ms preferred	G.1010 [23]
	<400 ms limit	TS 22.105 [24]
	<150 ms	TR-126 [25]
Videophone	<150 ms preferred	G.1010 [23]
	<400 ms limit	TS 22.105 [24]
Interactive games	<200 ms	G.1010 [23]
		TR-126 [25]
	<75 ms preferred	TS 22.105 [24]
	<50 ms (objective)	TR-126 [25]
Web browsing	<2 s/page preferred	G.1010 [23],
	<4 s/page acceptable	TR-126 [25]
	<4 s/page	TS 22.105 [24]

Table 1.2: Canada peak and off-peak latency comparison [22]

Technology	Peak Hour Latency	Off-Peak Hour Latency
Cable/HFC	14 msec	13 msec
DSL	12 msec	12 msec
FTTH	4 msec	4 msec

much better user experience. Table 1.1 shows the one-way delay recommended by the ITU, the 3rd Generation Partnership Project (3GPP), and the Broadband Forum for different applications [21]. Table 1.2 shows peak and off-peak latency for different transmission technologies from volunteers located within a 150km radius of Halifax, Montreal, Toronto, and Vancouver [22]. Therefore for very long distant transmission, like intercontinental transmission, packet latency can play a critical role for live video streaming.

Network traffic is usually protected from packet loss by TCP/IP (Transmission Control Protocol over Internet Protocol). TCP/IP establishes the communication

channel and continuous handshaking between the transmitter and the receiver to guarantee the delivery of data over the internet. The receiving terminal acknowledges all received packets and requests any lost packets be re-sent. Requiring receipt acknowledgments for every packet dramatically reduces the bandwidth efficiency. The TCP/IP protocol is consequently not suited for real-time video streams. Communication over the TCP/IP protocol is highly inefficient, especially over long distances. In addition, buffering happens in every router between the sender and the receiver, which introduces enormous transmission delays. Therefore, low latency streaming is not feasible using TCP-based protocols like MPEG-DASH.

In contrast to TCP, UDP (User Datagram Protocol) is a connectionless protocol, lightweight, and fast, but it discards erroneous packets. It is compatible with packet broadcasts and multicasting, and its lack of retransmission delays makes it suitable for real time applications such as live video streaming.

There are three methods in communication systems to avoid or correct packet failure: Forward Error Correction (FEC), Automatic Repeat reQuest (ARQ), and Error Resilient Coding (ERC) [26].

Forward Error Correction adds redundant data to the original data in the transmitter. It is not suited for a communication channel with strong noise. Automatic Repeat Request is another error control and packet recovery approach for data transmission that outperforms live video streaming or very large file transmission. The latency added by ARQ depends on the retransmission process or round-trip time between the transmitter and receiver. The ERC approach (like Multiple description coding (MDC)) is resilient against packet corruption or noise by adding redundant bits before transmission.

The multiple description coding (MDC) method is our method of choice due to its suitability for noisy channels. MDC avoids packet failure because it creates multiple complementary and separately-decodable descriptions [27]. In the next chapter, we will describe FEC, ARQ, and different methods of ERC to explain why MDC is our method of choice in this thesis.

1.3 Problem Statement

With the recent emergence of 3D TV technologies, video service providers now have an opportunity to stream live 3D videos over networks. 3D video streaming, like

sport streaming which is popular with audiences nowadays, transmits massive data volumes, usually over long distance networks. Such transmissions require significantly increased bandwidth; with a lower latency compared to the normal data transmission applications. Network congestion and packet loss is more probable for such an application and needs to be mitigated.

For such video applications, retransmission is unacceptable or infeasible due to the long Round Trip Time (RTT) experienced on long distance networks. If the loss rate is greater than the value that the FEC scheme is designed initially, recovery of lost packets is not possible.

1.4 Research direction and contributions

As argued in Section 1.3, the increased demand for resources (including higher bandwidth and lower latency) in 3D sport streaming applications results in higher packet loss. Current methods of 3D video compression and streaming rely on the former 2D video compression and streaming methods, which are not efficient.

MDC as a major error resilient method of video streaming is the method of choice in this thesis. It fixes the errors that occur in one description by considering other error free descriptions. MDC is not without its drawbacks. It contributes to encoding inefficiency, as it decreases dependency between data in each description [27]. Temporal MDC, the most common MDC method, may not be suitable for live 3D sport event streaming, which requires low latency, especially in error-prone networks and over long distances. Therefore, to apply an MDC method to a 3D video, we need to first look for the best-suited MDC type for live stream communication such as sport event streaming; second, to increase MDC performance regarding what type of data needs to be added in each description. Also, as human eyes are more sensitive to objects than pixels, is it best to divide the video data into descriptions without considering the importance of a scene's objects?

Therefore, this thesis

- describes how 3D videos are captured, streamed, and displayed;
- examines various methods of video streaming to find the most appropriate approach to tackle the packet loss that happens in error prone networks;

- surveys different methods of layered video coding and multiple description coding;
- introduces a new MDC method considering important objects of the scene and examines its performance.

The introduced method assigns more bandwidth in each description to the important objects of the scene. This way the coding efficiency does not drop significantly because the enhanced data in each description is highly dependent on the original data of that description. It is efficiently compressed using differential pulse code modulation. Moreover, important objects are decoded with a better quality, improving users' experience, as human eyes are more sensitive to objects than pixels.

Generally, when a video is recorded, there are one or more important objects in the foreground in addition to usually not important objects located in the background. To extract the important objects, this thesis uses the fact that the depth map image of a 3D video contains low frequency contents since the depth information of pixels of an object are very close to each other. We introduce a simple algorithm to extract those objects first and then add important objects' color information to each description.

1.5 Publications

The following is a list of publications generated during the course of the Ph.D. program.

1. **E. Rahimi**, C. Joslin, Video Spatiotemporal Multiple Description Coding Considering Region of Interest, In Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, 474-481, 2020 , Valletta, Malta.
2. **E. Rahimi**, C. Joslin, Reliable 3D video streaming considering region of interest, In EURASIP Journal on Image and Video Processing volume 2018, Article number: 43, 2018.
3. J. McAvoy, **E. Rahimi**, C. Joslin, Low Complex Image Resizing Algorithm using Fixed-point Integer Transformation,In Proceedings of the 13th International

Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, 143-149, 2018 , Funchal, Madeira, Portugal.

4. **E. Rahimi**, C. Joslin, 3D Video Multiple Description Coding Considering Region of Interest, In Proceedings of the 12th International Conference on Computer Vision Theory and Application, 2017, Porto, Portugal.

1.6 Thesis Organization

The rest of this thesis is organized as follows:

- Chapter 2 introduces 3D/Multiview video capturing, representing, and streaming. It provides a background on the error robust method of video streaming focussed on multiple description coding. The chapter concludes by representing the state of the art.
- Chapter 3 more fully introduces the method of choice. The implementation algorithm will be explained in this chapter.
- Chapter 4 presents the experiments and results.
- Chapter 5 describes a summary of this work, along with proposed future research directions.

Chapter 2

Background

2.1 3D/Multiview Video Capturing, Representing, and streaming

The beginning of 3D TV goes back to 1838, when the stereoscope was invented by an English scientist, Sir Charles Wheatstone. He showed that when two pictures are viewed stereoscopically, the human brain combines them to produce 3D depth perception [28]. Today, with the help of recent technological developments, including 3D displays, real-time video processing capabilities, and telecommunication services such as 5G, we are now facilitating access to 3D/multiview applications in audiences' homes easily.

Before talking about various 3D/multiview streaming methods, in this chapter, we are going to explain how 3D videos are captured, displayed, and transmitted.

2.1.1 3D Video perception

The human eyes in the horizontal plane are located 65 mm apart, approximately [29]. Therefore, scenes are captured by our eyes from each eye's different perspective, resulting in depth information of the scene also known as retinal disparity [29]. Typically, depth information gained by the human eye is called depth cue.

Depth cues can be obtained from different sources such as:

- the focal depth cue can be measured based on the refractive power of the human eyes to produce a clear image at different distances from it;
- the binocular depth cue comes from retinal disparity;

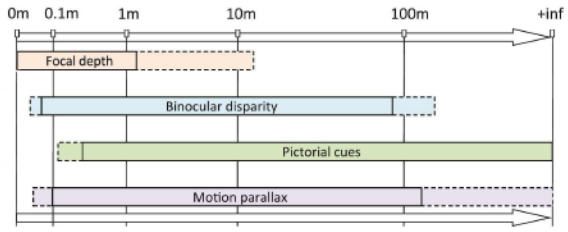


Figure 2.1: Importance of the depth cues based on the distance of the eyes from an object.

- pictorial depth cues such as shadows, texture scaling, etc. are highlighted for large distances;
- the motion parallax cue, also known as the head parallax is caused by head motion and describes a situation in which closer objects to viewer's eyes seem to move faster than further objects.

Although depth cues are obtained from several resources, it is the binocular disparity depth cue that is used with stereoscopic videos. As can be seen in Figure 2.1, the importance of depth cues is varied for different distances [30].

2.1.2 3D/Multiview Video Representation

As described by MPEG-3DAV, 3D videos can be captured and represented as stereoscopic video or multiview video [2, 17]. It is worth mentioning that 3D videos can also be captured and displayed by a hybrid method. For example, 3D video can be captured by an array of cameras (multiview method) but represented as stereoscopic video. This method, called "Free viewpoint" by Tanimoto [31], gains from the motion parallax depth cue, which could be obtained by the multiview method, while less bandwidth is required to deliver the 3D video. The following is a brief description of different 3D video representations:

- Multiview video representation can be considered as the general case of all 3D video representations. With this type of representation, an object is viewed from different view directions, although the display view direction is not limited to captured views directions, and novel views can be made via interpolation of

the real camera view directions. In order to be used commonly, this type of representation needs to be delayed until the necessary infrastructure and more effective compression encoders are available, as multiview video requires more storage space, bandwidth availability, and processing power. Free viewpoint television (FTV) and surveillance are some typical applications of this type of representation [32].

- With stereoscopic representation, a video captures only from two adjacent views of a scene like human eyes. On the other hand, stereoscopic video is a special case of multiview that needs fewer resources of storage space, bandwidth availability, and processing power. Therefore stereoscopic videos are more commonplace [32].

- To gain the advantage of multiview video representation, i.e., having an arbitrary view and the advantage of stereoscopic video representation, i.e., requiring less bandwidth availability, the scene can be captured by an array of cameras as a multiview video but the two views closer to the position of the observer's head would be chosen and streamed toward the receiver. The decoder uses the received views to reconstruct the stereoscopic video for the user. Therefore for this type of representation, feedback needs to be sent by the receiver for each time slot to let the encoder know which views need to be chosen for the next time slot. This way, a motion parallax depth cue, which is ignored by the basic stereoscopic representation, would be exploited to produce a more effective 3D video representation. Such an approach, called "Free viewpoint," also can interpolate for a virtual view to find the best view close to the observer's head position [31, 33].

The stereoscopic video representation with its simple camera adjustments, lower space disk, and bandwidth requirements make it the most popular of the three 3D video representations. To produce stereoscopic video, one of the following methods is used:

- **2D to 3D conversion:** This method produces the depth information from an existent 2D video using different algorithms [34–44]. As an example of such algorithms, Philips announced the WOWvx BlueBox that provides a 3D clip from a 2D video in a semiautomatic algorithm [44]. One important concern

regarding the 2D to 3D conversion method is to have an automatic real time conversion from 2D video into 3D video. Sisi et al. have clarified the differences between off-line and on-line algorithms of 2D to 3D conversion and then explained the principles required for the online algorithms [42]. Harman et al. described an algorithm of a 2D to 3D conversion method in which the depth information can be derived from the motion data, although it is not applicable for all 2D videos [36]. In work done by Konrad et al. [43], two automatic types of conversion, i.e., point mapping and depth map estimation, are developed for some special videos. Another approach, presented by Lai et al. in [41], estimates depth information from a 2D video considering motion information, linear perspective, and texture characteristics. More efficiently, Chang et al. have introduced an adaptive approach to refine a depth map obtained by a 2D to 3D conversion algorithm [39].

- **Dual camera configuration:** The basic and economical method of capturing a 3D scene is recording it with a stereo camera because there is no need for preprocessing in the video player to display the 3D video. Depth information estimated with this method can vary according to the shooting parameters, such as base distance, convergence distance, and focal length. Base distance means the horizontal distance between the two cameras, and convergence distance describes the distance between cameras and the intersection point of both views' axes. Generally, there are two types of camera configuration for capturing a 3D video with a stereo camera: parallel camera configuration and toed-in camera configuration [2]. Both types of configuration produce left and right views, and no more processes are required to display the 3D video in the video player. Between these two types of configuration, parallel camera configuration outperforms in terms of geometrical distortion [45]. For example, the keystone effect (also known as tombstone effect) is caused when the view axes are not perpendicular to the projection plane, which is more likely in the toed-in camera configuration. The keystone effect distorts the image dimensions so that a square looks like a trapezoid. In addition, the depth plane curvature is usually prevented in the parallel configuration. Depth plane curvature distorts an image so that objects placed in the center of the image seem closer to the observer's eye compared to the objects located in the corners of the image [46].

- **3D/Depth-range cameras:** Another method of producing 3D video is capturing images with a depth-range camera. This method, which is used by Zcam^{TM} [47], Axi-vision [48], or Kinect [49] captures 2D video along with an associated sequence of depth information obtained via light or ultrasound pulse. Therefore to display 3D video in the player, left and right views need to be generated from a 2D image plus depth map image.

As mentioned earlier, this thesis mainly focuses on the color image plus depth map image representation, which can be derived from either left and right views or depth-ranges cameras. Clearly, the depth map image mainly contains depth information of the scene objects, and due to the nature of real objects, such information rarely contains high-frequency components [2]. Therefore depth information can be compressed more effectively than color image, and consequently bandwidth and memory space will be saved much more compared to the left and right views [2, 50]. Also, 3D scene reconstruction can be adapted according to the users' preferences. For example, a user can watch a video from a novel view different from the captured views because the stereoscopic representation is not limited to existent views, as 3D image warping is done by the receiver [2].

It is worth mentioning the depth map image using by this thesis has a similar spatio-temporal resolution as the color image sequence. The depth value for each pixel is stored in 8 bit, from 0 to 255.The value 0 specifies the further value and the value 255 specifies the closest value to camera. Considering a linear quantization of depth, the real depth value can be calculated using Equation (2.1) [2].

$$Z = Z_{far} + d(\frac{Z_{near} - Z_{far}}{255}), d \in [0, \ldots, 255], \tag{2.1}$$

where d specifies the respective depth value as shown in Equation (2.1). In general the colorful 2D frame is called as color image or Texture image and the gray scale depth frame is called as depth map image.

2.1.3 3D Video Streaming

To stream a 3D video, it needs to be encoded differently from 2D videos because 3D videos include disparity redundancies in either form of representations mentioned in the previous section. Because of the dependency between the two views, the 3D

Figure 2.2: sample color image plus depth map image.

encoding algorithm should remove inter-view redundancies in addition to the spatial and temporal redundancies. There are several methods to stream stereoscopic video using available encoding standards. These standards include simulcast configuration, frame packing configuration, frame compatible, 3D video layered coding, 3D video multiple description coding, etc. The standards are explained in the following:

- **Simulcast configuration:** Simulcast configuration describes a method in which texture image and depth map image are encoded separately using two H.265 encoders or any former video coding standard. With the simulcast configuration, two streams will be generated (see Fig. 2.3), and then they are multiplexed to make one stream. Therefore, the simulcast configuration can be treated as a post-processing configuration, and a demultiplexer will be required in the decoder to split the color image stream and depth map image stream. The biggest advantage of the simulcast configuration is its simple structure; however, it is less efficient, as both views are encoded separately and disparity redundancies are not removed [2].

- **Frame packing configuration:** Frame packing configuration packs both views' frames first, and then the combined frames are encoded using one HEVC encoder. This type of configuration can be treated as a preprocessing configuration, and left and right views' frames can be combined by many different methods. A simple combination can be the side by side configuration, which provides a frame with twice the resolution (see Fig. 2.4a).

  Another scheme, known as the top-bottom scheme (see Fig. 2.4b), locates one

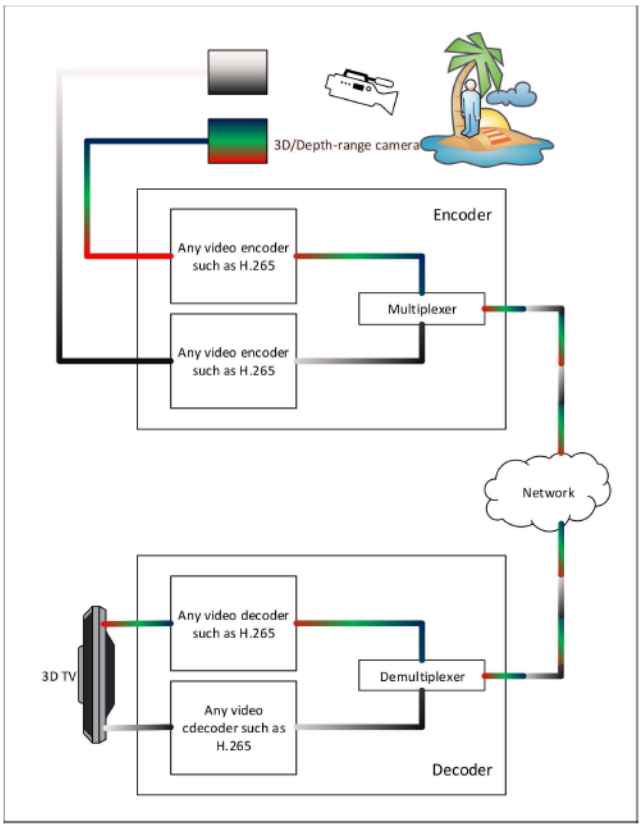
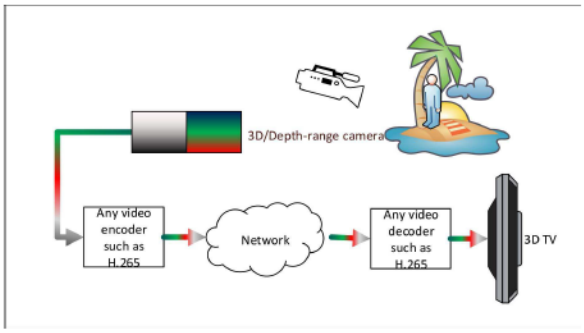


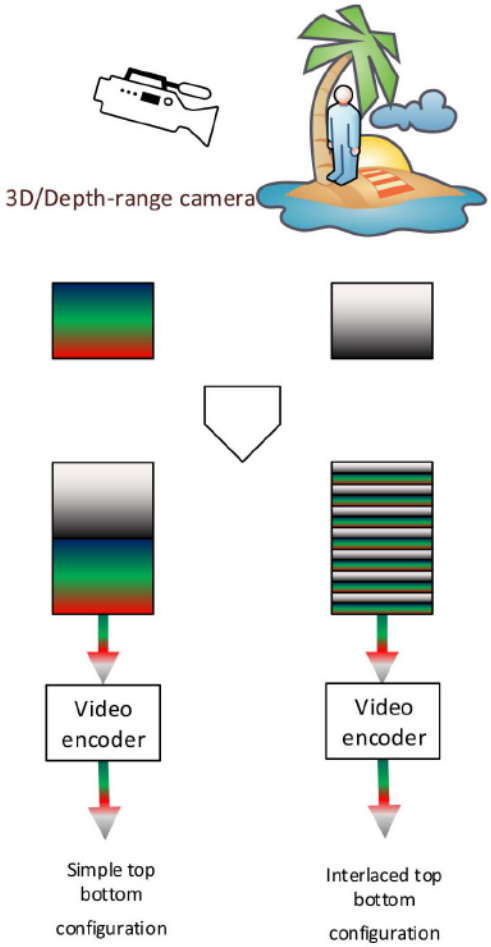
Figure 2.3: Simulcast configuration: gray-scale box represents the depth map image and the color box represents the color image.

view above the other one either completely or line by line (interlaced top-bottom approach). Similar to a simulcast approach, a frame packing configuration is also less efficient because of not considering inter-view redundancies. However, only one stream will be generated this way, and there is no need for a multiplexer [2].

- **Frame compatible:** The 3D frame compatible approach follows similar rules to the frame packing approach. The only difference is that the frame compatible approach generates a packed frame with equal resolution to original left or right views. Clearly, this can be done via downsampling of the views (see Fig. 2.5). Just like previous configuration, this category can be also done as side by side configuration, top bottom configuration, or top bottom interlaced configuration [2].



(a) side by side configuration.



(b) Top-bottom configuration: simple top-bottom approach (left) and interlaced top-bottom approach (right).

Figure 2.4: Frame packing configuration.

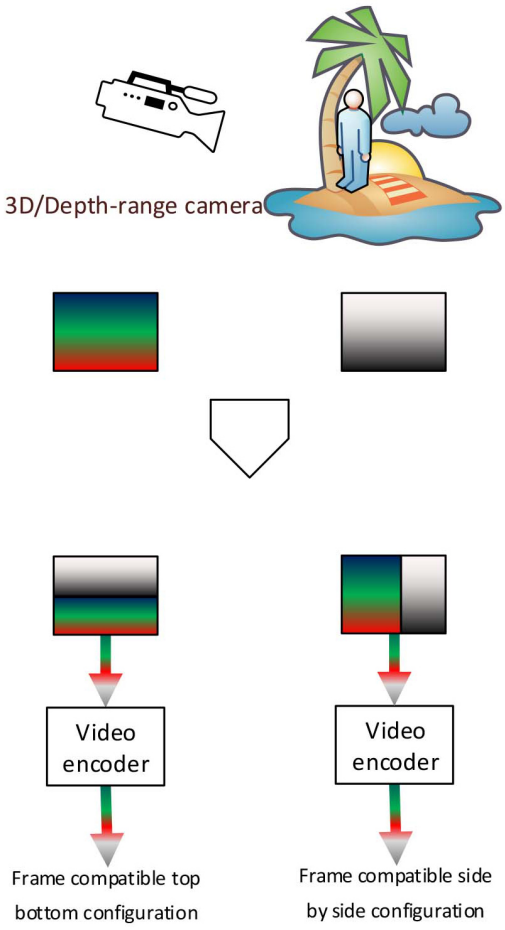


Figure 2.5: 3D frame compatible configuration.

2.2 Error robust method of video streaming

Network traffic is error free if packets are transmitted using the TCP/IP protocol. As shown in Figure 2.6, using the TCP/IP protocol, the receiver must acknowledge all received packets and inform the transmitter if a packet is lost. Such a process may drop bandwidth efficiency dramatically and make the TCP protocol not best suited for video streaming. Compared to TCP, UDP is a lightweight protocol that does not require acknowledgments. In other words, UDP does not do error correction and assumes it is performed in the application, to avoid the overhead of the handshaking processing at the network interface level. It is worth mentioning that IPTV services that use UDP protocol have approximately 123 million subscribers around the world and their subscribership are rising at a rate of 12 percent per year [51].

Using UDP, packet loss needs to be corrected using forward error correction (FEC), automatic repeat request (ARQ), or error resilient coding (ERC) [27].

2.2.1 Forward Error Correction

Forward error correction adds redundant data to the original data in the transmitter and uses the redundancy at the receiver to detect and/or correct transmission errors. For example, it duplicates raw data so that if one fails, the original data can be recovered; however, if both data are lost there is no chance to recover the original data.

For example, an FEC encoder with a linear block code with the rate of $\frac{n}{k}$ converts n message bits to k transmission bits and transmits through the channel. Such code can detect all combinations of $d_{min}-1$ or fewer bits errors (d_{min} is the minimum Hamming distance between two words in the code alphabet). Also, a block code with minimum Hamming distance d_{min} guarantees correcting all patterns of $t = \left\lfloor \frac{d_{min}-1}{2} \right\rfloor$ or fewer errors. For instance, a simple 3 repetition code is able to detect error up to 2 bits and recover the error only if one bit fails;

Therefore FEC scheme is a channel coding technique which is designed and implemented at the transmitter to detect and/or correct specific amount of bit error at the receiver; if more data bits are lost, there is no chance at receiver to recover the original data error-free. It is possible to over-design the code to deal with more errors however efficiency of such scheme is reduced. Since channel condition is fluctuating

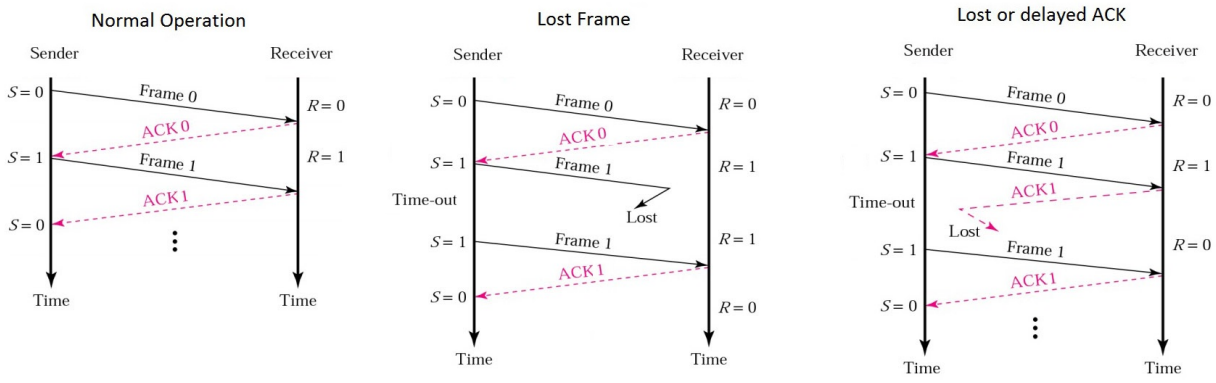


Figure 2.6: Normal and lossy TCP systems transmission.

and loss rate is not usually static, FEC protection can not always produce a good approach.

Also, FEC is not well matched with the burst errors that happen in public internet. FEC is typically appropriate for networks with low and consistent packet loss or when establishing a relationship between the transmitter and receiver is not possible, like satellite communication [53, 54].

In other word, FEC increases bandwidth overhead, as it generates redundant data to protect against networks with "known lossiness. FEC also adds latency in the range of 100 to 500 milliseconds, as it must perform the initial calculations in the transmitter and provide for the recovery calculations at the receiver. Another drawback of using FEC is the fact that to avoid over-provisioning overhead (extra latency and bandwidth), the operator needs to know the packet loss pattern, which is not possible for internet connections and video streaming [55, 56].

2.2.2 Automatic Repeat Request

In the ARQ method, the receiver requests retransmission if a packet is missed, and the transmitter resends the missed packet only if it receives the request (see Figure 2.7). Compared to the TCP/IP protocol, ARQ only acknowledges if a packet is missing and

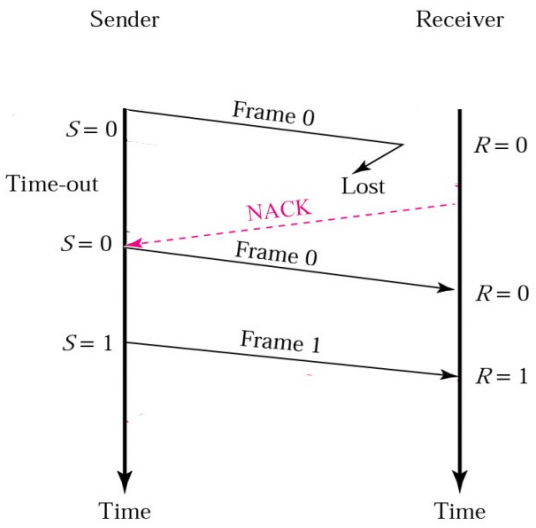


Figure 2.7: ARQ system's retransmission.

does not acknowledge the receipt of every packet. Therefore, it avoids the massive communications that make TCP highly latent and inefficient with bandwidth.

The bandwidth overhead of the ARQ method is proportional to the network packet loss and does not have a fixed breakable threshold level like FEC. Such behavior makes it more appropriate for unpredictable networks such as public internet [52, 54]. However ARQ is more appropriate for video streaming. It requires a duplex channel and cannot be used in a unidirectional channel. Also, it is not efficient because of the ARQ system's retransmission. Moreover, the latency added by the ARQ method depends on the retransmission or round-trip time between the transmitter and receiver. With more distance between the transmitter and receiver, more latency must be accommodated through larger buffers [52].

2.2.3 Error Resilient Coding

The ERC approach provides resiliency against packet corruption or noise features by adding redundancy bits before transmission. There are a number of methods which redundancy can be introduced to the stream at the source coding, including reversible variable length coding (RVLC), intra refreshment, flexible macroblock ordering (FMO), layered coding (LC), multiple description coding (MDC), etc.

2.2.3.1 Reversible Variable Length Coding

RVLC is referred to the VLC codes which can be instantaneously decoded in both direction. In VLC, if bit errors happen in the middle of a codeword, the codeword and all following codewords (even if they are received correctly) are undecodable and the VLC bitstream is corrupted; because the decoder fails to locate correct boundaries of VLCs due to codeword synchronization problem.

One simple solution is inserting resynchronization markers periodically however it will reduce the coding efficiency. Therefore, if an error occurs, the decoder discards all the bits until the next resynchronization marker.

Another fix for VLC synchronization loss is RVLC in which the decoder can not only decode bits after a resynchronization codeword, but also decode the bits before the next resynchronization codeword in backward direction. In other words RVLC codewords can be parsed in both the forward and backward direction, making it possible to recover more data from a corrupted data stream. Although RVLC helps the decoder to find more errors and provides more information on the position of the errors compared to VLC, the RVLC used in H.263 lacks coding efficiency [57].

2.2.3.2 Flexible Macroblock Ordering

Flexible macro-block ordering (FMO) is one of the error resilience tools included in H.264/AVC however it is usually more beneficial for the channels with lower bit error. FMO assigns each macroblock freely to a specific slice group using a macroblock allocation map (MBAmap). It is possible to have up to eight slice groups in one picture and within a slice group, macroblocks are coded in default scan order. Within a certain slice group, the macroblocks can be grouped into several slices. The scenario that FMO is not used by the encoder is when there is only one slice group within a picture. Slice groups can be done in a way that, each lost macroblock may be surrounded by macroblocks of other slice groups. In that case, the lost macroblock can be reconstructed in a very effective way using interpolation based on surrounding (available) sample values [58].

This technique is well suited to real-time, ultra-low-delay applications however it is not beneficial for the unequal error protected transmission channels. It also adds unacceptable bitrate overhead. The coding performance could also drop due to a smaller or dispersed search space for inter prediction [58, 59].

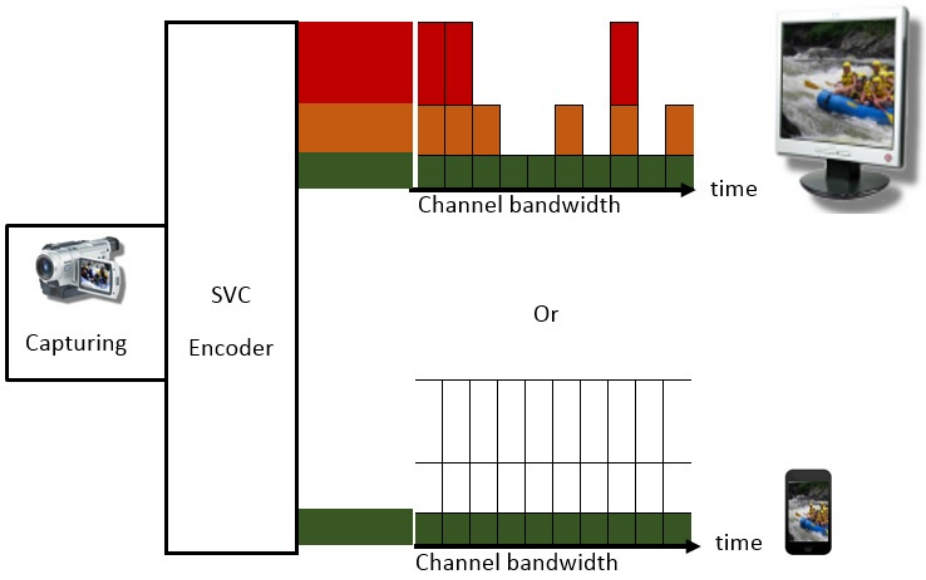


Figure 2.8: Scalable video coding: Upper case displays the situation that a client requests the full HD resolution but cannot display the full HD sometimes because of the channel fluctuation.
The lower case describes a situation where a client's device cannot display full HD video. Therefore client's device ignores the enhancement layers and only decodes the base layer.

2.2.3.3 Layered Coding

In addition to quite a few situations on the client-side, the network capacity and channel state are also changing continuously, resulting in unguaranteed bandwidth allocation.

Such different receiver device capabilities and channel fluctuations have motivated researchers to look for a flexible video coding method called layered coding (LC) or scalable video coding (SVC).

Scalable video coding encodes a video into one base layer (BL), and several enhancement layers (EL) based on the frame rates, resolutions, or qualities [60, 61]. In other words, layered encoded bitstream is decodable either wholly or partially according to different situations of multimedia platforms, communication channels, and video players (see Fig. 2.8).

In such a flexible coding scheme, the encoder streams the upper layers only if there

are still available resources (for example, bandwidth) after transmitting the lower layers. This method might lose its efficiency if facing a very noisy communication channel because the base layer may fail to decode successfully [62–65]. In this case, it is better to transmit BL using error-supported methods, such as multiple description coding.

The last scalable video coding standard, Scalabilty extension of HEVC (SHVC) [66] and the scalable extension of the former video coding standard, H.264/SVC [61] provide scalability tools to stream video using layered coding, which is explained in the following. These scalability tools allow the encoder to provide three types of scalability: temporal, spatial and quality scalability [30, 67].

The layered coding method, which can be done by the scalable extension of video coding standards, such as SHVC [68], SVC [61, 69], or MPEG-2 TS [16], creates one bitstream in one base layer and several enhancement layers but removes interlayer redundancies. This method also has the capability of streaming the video data dynamically. An essential criterion of this method is backward compatibility to provide service for those devices that do not support the layered coding method. In other words, customers can see 3D video, full HD 2D video, or 2D video with a lower resolution depending on the receiver device's features.

In layered coding, the layers are not separately decodable, resulting in performance dependency upon lower layers to be without error. Therefore, layered coding is less advantageous for error-prone environments.

2.2.3.4 Multiple Description Coding

In contrast to hierarchy substreams of layered coding, multiple description coding (MDC) partitions a single video stream into two or more substreams (referred to as descriptions) in parallel [27]. Then packets of each description are streamed over various paths to benefit the MDC technique's error resiliency aspect. The receiver can decode the video stream using any description; however, using more descriptions improves decoded video quality.

To clarify how MDC is tolerant of packet failure, suppose packets drop as they traverse the network due to node congestion, packet corruption, and significant packet delay (expected in best-effort networks such as the internet). The probability of the event that packets regarding all descriptions for the same video data are dropped is very low because packets pass over different paths. In other words, in case of an error

in one description, other descriptions are available to reconstruct the video at the receiver. Therefore only quality fluctuates and the entire video will not be affected.

Besides its packet failure robustness, MDC also allows for rate-adaptive streaming. In contrast to MDC's increased fault tolerance, it decreases the total compression ratio because each description needs to include extra information as the header; such a cost of coding inefficiency is unavoidable, as each description needs to be separately decodable.

The above argument is not the only reason for dropping the coding efficiency. MDC lowers the compression ratio, as each description's data is not as dependent as it was in the original video stream; so, the differential pulse code modulation (DPCM) technique used by the encoder is not as efficient as it was before.

The encoder needs to pack the highly correlated data in one description to keep the coding efficiency; however, it needs to distribute highly correlated data between descriptions to keep the estimation power high for the missed description from other descriptions. Therefore, there is an error resiliency-coding efficiency trade-off problem. For example, temporal MDC usually creates only two descriptions because it keeps the data dependency high enough; therefore, the coding efficiency is higher [63, 70].

Fig. 2.9 shows the MDC technique. As can be seen, there are two types of decoder, called the central decoder and the side decoder. If all descriptions are received, the central decoder takes care of the decoding process; otherwise, the side decoder is responsible for decoding. However, it may produce some distortion. Although the central decoder provides better quality, the side decoder offers a better quality of experience if the current rate is not enough to support all descriptions, or transmission channels are so noisy that some descriptions may be received unsuccessfully.

There are several types of MDC, as the data distribution between descriptions can be chosen from various domains, such as the spatial, temporal, or frequency domains.

2.2.3.4.1 Spatial Multiple Description Coding

Spatial multiple description coding distributes adjacent pixels of a frame to different descriptions. A polyphase subsampling algorithm in the encoder produces lower-resolution images called "sub-image," as shown in Fig. 2.10. Based on the availability of the descriptions in the receiver, the decoder chooses the central or side decoder to reconstruct the video.

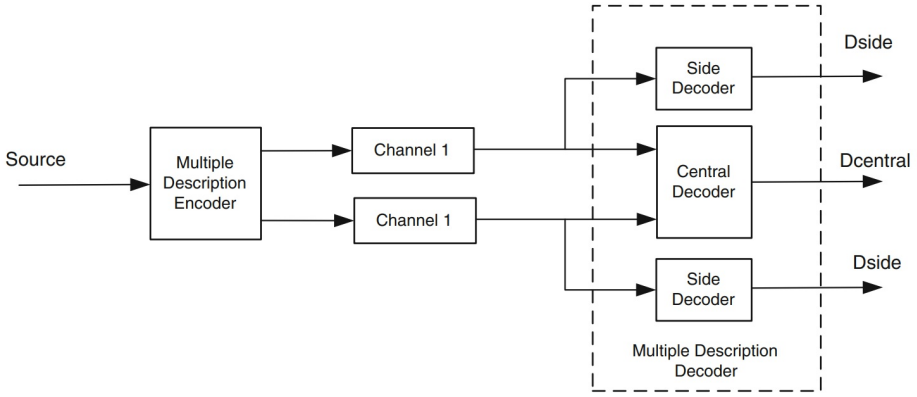


Figure 2.9: Multiple description coding. Reprinted by permission from Springer: Springer Multimedia Systems [27] ©2014.

Although such a type of multiple description coding is beneficial in noisy environments, there is no precise adjustment tool over the redundancy to control the side quality. Shirani et al. introduced the zero-padding approach (to add the adjustment tool) by which the encoder can set the amount of redundancy more precisely [71, 72]. The zero-padding method transforms each frame's pixel values to a discrete cosine transform (DCT) domain first. Next, it adds zeros to the transformed data (Fig. 2.11) and then retransforms back to the spatial domain. Afterward, the polyphase subsampler (PSS) partitions the new video frames' pixel data into multiple descriptions.

It is worth mentioning that adding zeros in the frequency domain is interpreted as upsampling in the spatial domain. Therefore, this approach adjusts redundancy as required and then creates MDC descriptions.

A more recent approach adds zeros in vertical and horizontal directions in the frequency domain as the transform coding for video is applied on a two-dimensional-block. However, Tillo and Olmo argued that it is more efficient to add zeros in one direction, the direction in which the image's pixel failure is hard to estimate [73].

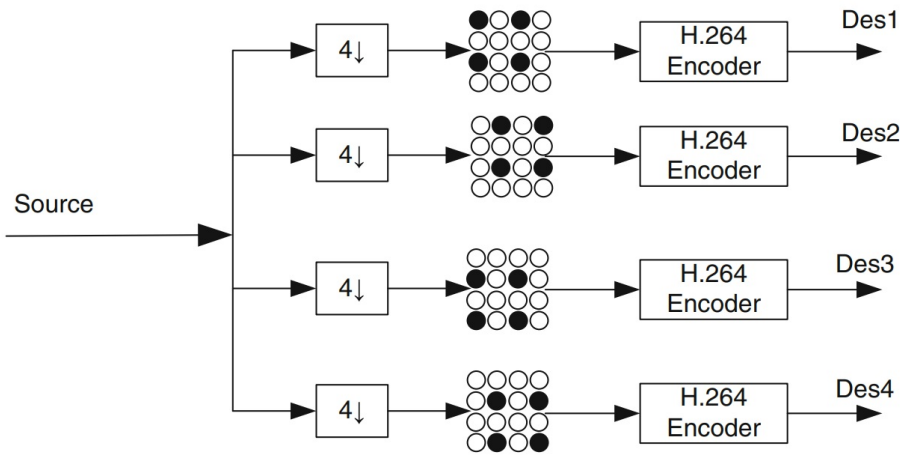
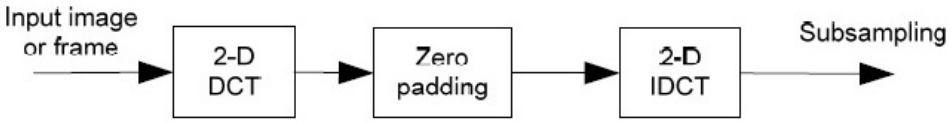
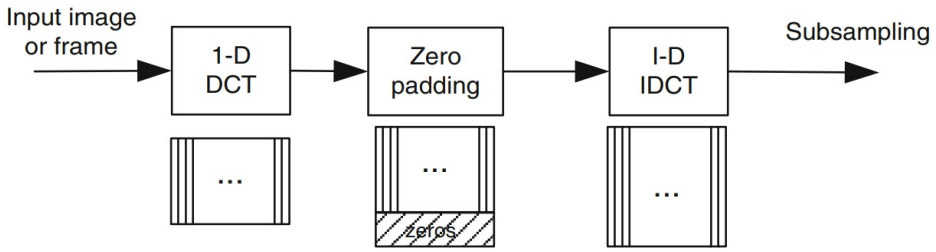


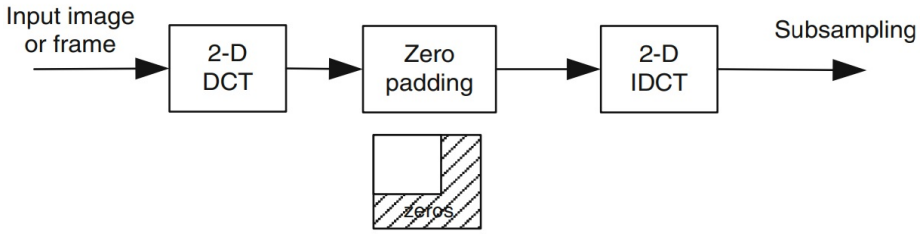
Figure 2.10: Basic spatial PSS multiple description coding. Reprinted by permission from Springer: Springer Multimedia Systems [27] ©2014.



(a) Zero padding process ©IEEE 2001.



(b) One-direction zero padding process. Reprinted by permission from Springer: Springer Multimedia Systems [27] ©2014.



(c) Two-direction zero padding process. Reprinted by permission from Springer: Springer Multimedia Systems [27] ©2014.

Figure 2.11: Zero padding approach presented in [27, 72].

Tillo and Olmo [73] also introduced another spatial MDC method called "the least predictable vector directional multiple descriptions coding." Their MDC method copies the least predictable part of the frame in all descriptions. The simulation result shows that this method improves the side quality compared to the previous simple PSS approach, although the new method provides more redundancy. They also argued that this approach is more complex, as it needs to detect the frame's least predictable data. The other issue that this method has is its restriction on the number of descriptions.

As argued earlier, MDC estimates the dropped description with the help of other error-free descriptions. One standard solution to calculate the lost pixels is a linear interpolation.

Padmanabhan et al. improved the accuracy of estimation by introducing a nonlinear spatial MDC method. This method uses an optimal filtering process to interpolate the missed pixels compared to the simple linear interpolation [74]. Fig. 2.12 shows how this method works for two scenarios: a) only one description is available, or b) two descriptions are used among four descriptions. According to Padmanabhan's work, the encoder designs a proper filter and sends filter coefficients along with the descriptions' data stream. This method's performance depends on the type of filter and how accurate it is. Also, the designed filter's complexity must not exceed the maximum allowed complexity for a specific application. For example, for a business video conference application in which no delay is allowed, this method may not be feasible if the increased complexity causes delay. According to the result shown in [74, 75], this method provides better performance than the zero-padding approach and the linear interpolation. Ates et al. combined the zero-padding approach and optimal filtering technique [75] and showed it has better performance in comparison to the result presented in [74].

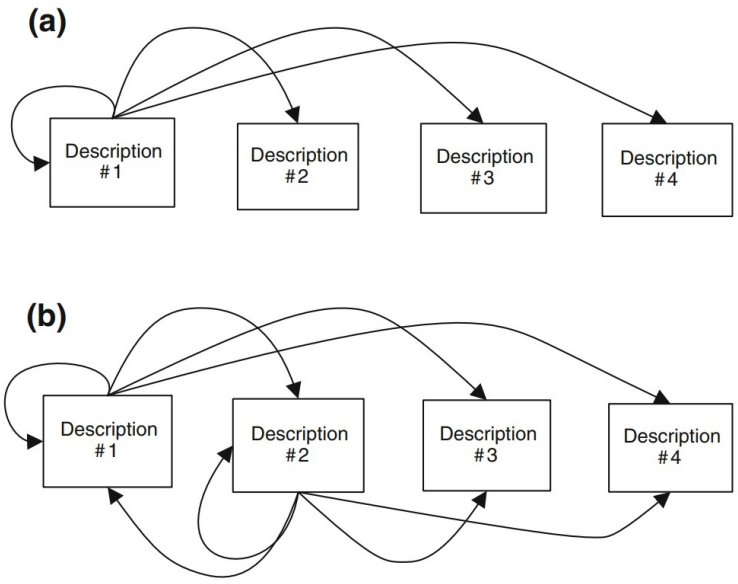


Figure 2.12: Optimal filtering multiple description coding: a) Situation that only first description is received. b) Situation that first and second description are received. Reprinted by permission from Springer: Springer Multimedia Systems [27] ©2014.

Shirani also presented a nonlinear PSS approach and analyzed its performance in case of missing one or more descriptions [76]. According to his work, the frame's interesting area (called the region of interest (ROI)) is sampled at a greater rate than areas judged not important, based on an exponential equation. In other words, descriptions include more information regarding the ROI, enhancing the side quality in the side decoder. For example, image b in Fig. 2.13 is a nonlinear transformation of image a to highlight the ROI (face). Also, the image d, which has been reconstructed from only one description, has an acceptable quality in comparison to image c, which has been decoded by the central decoder (having four descriptions). It is worth mentioning that since human eyes are more sensitive to objects than pixels, this method can also provide a better performance from the point of subjective assessment; however, this work did not discuss how the ROI can be found. This problem for the applications involving fast video content is more sensible.

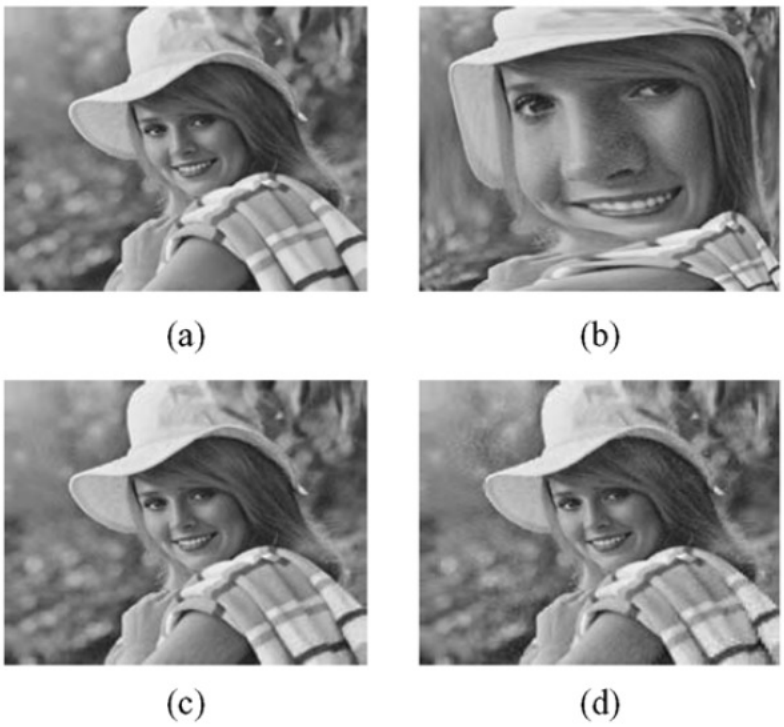
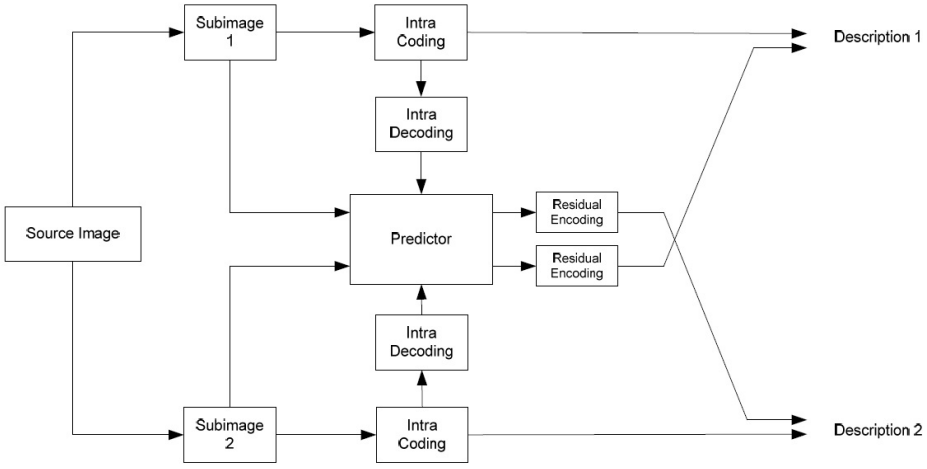
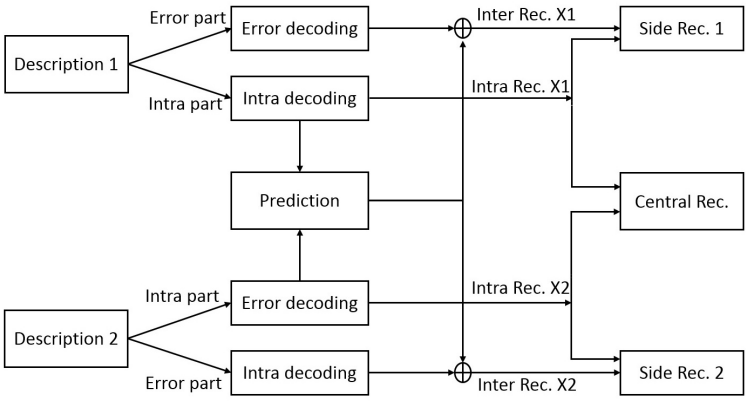


Figure 2.13: Content based multiple description coding: a) orginal picture. b) non linear scaling of original picture. c) decoded by central decoder. d) decoded by side decoder (only one description) [76] ©IEEE 2006.

In another spatial MDC method, introduced by Wang et al., first, the encoder estimates each sub-image from others, then it calculates the residual signal and streams it along with each sub-image as side information [77, 78]. Fig. 2.14 shows how this method reconstructs a missed description from the available ones. This figure assumes only two descriptions are available. The performance of this method depends on the quantization pace of the residual signal. In other words, the quantization parameter is the adjustment tool to control the redundancy and enhance or reduce the quality of the reconstructed description.



(a) encoder



(b) decoder

Figure 2.14: Predictive multiple description coding [77] ©IEEE 2006.

Zhu et al. introduced a new polyphase down-sampling based MDC, called NPDS-MD. The framework of the proposed NPDS-MDC method is shown in Figure 2.15. This method consists of three major steps [79]:

- first, pixels are down sampled according to quincunx pattern,
- second, in addition to each down-sampled description, the descriptions are transformed and side information is produced to send along with description.
- third, an error compensation algorithm is applied to each description to mitigate the compression distortion.

Their simulation results shows the superior performance, especially at high bit rates; however at the expense of more bandwidth and complexity. Moreover, this method only produces two descriptions and therefore is not suited when more description are needed. Furthermore, its complexity limits its application to image transmission and not live video streaming.

The other type of spatial MDC method is multi-rate MDC, presented by Jiang and Ortega. Fig. 2.16 shows the structure of such a spatial multi-rate MDC approach [80]. According to Jiang and Ortega's work, the low-quality version of the other descriptions is added to each description to help the side decoder enhance its estimation power.

The mentioned spatial MDC approaches create descriptions symmetrically, causing the same side qualities for all side decoders. However, there are also MDC approaches creating nonsymmetric descriptions and, therefore, varying side quality. Although the reconstructed video in the central decoders has better quality, such algorithms are not usually optimum for applications dealing with low noise levels, as more redundancy is generated by this method.

The MDC method presented in [81] by Zhu and Liu is from this nonsymmetric category. Their proposed method creates descriptions using various references for the DPCM algorithm or different directions of the motion estimation used by the hybrid video encoder. Two descriptions are made based on different quantization paces and a separate reference for the motion prediction algorithm. The central decoder calculates the weighted summation of these two descriptions to produce a lower error. The encoder generates the weights used for the summation in the central decoder via estimation theory. Then it quantizes them and sends them as side information.

Another nonsymmetric MDC algorithm introduced by Zhao et al. changes the block borders for different descriptions. Video coding standards perform motion estimation and compensation in the video coding in a block-wise manner. Therefore, using various blocks for each description yields a different residual signal, and consequently, a different quantization pace is required. The central decoder uses the difference between two descriptions to decrease the reconstruction error [82]

2.2.3.4.2 Temporal Multiple Description Coding

Different descriptions in a temporal MDC method are usually created by assigning

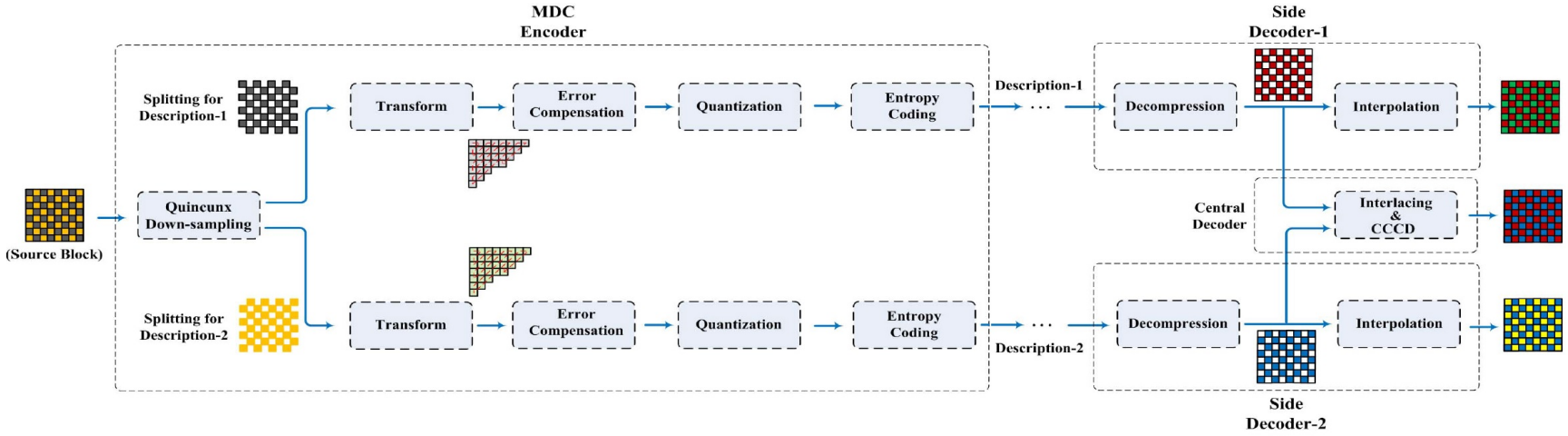


Figure 2.15: NPDS-MDC block diagram [79] ©IEEE 2020.

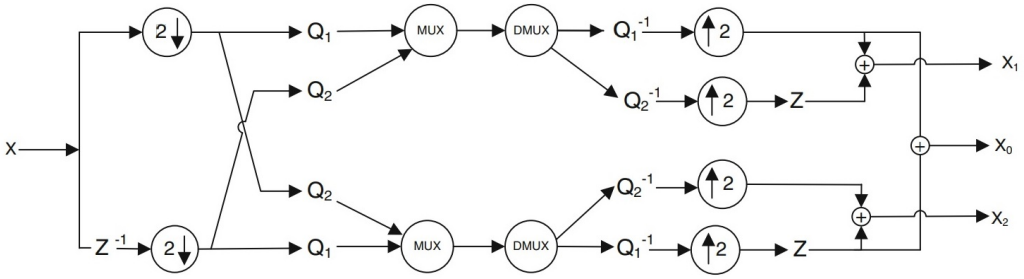


Figure 2.16: Spatial multirate multiple description coding. Reprinted by permission from Springer: Springer Multimedia Systems [27] ©2014.

different frames to each description. For example, one description includes only the odd frames while the other description includes only the even frames (Fig. 2.18).

For example, in [83] Dinh introduced a new temporal MDC approach benefits of the new H.266 Versatile video coding standard (the successor to High Efficiency Video Coding, HEVC, which aims to make 4K broadcast and streaming commercially feasible [84, 85]) and also distributed video coding (DVC) standards [86]. As can be seen in Figure 2.17a, the encoder encodes the source video sequence into two odd and even subsequences and then transmit these descriptions to the receiver. At the receiver, the decoder uses the WynerZiv (WZ) coding, introduced in the DVC, to provide a high image quality for the video sequence. This way, the redundant data can be effectively controlled based on the WZ coding scheme.

There are several methods to estimate missed frames using the available frames, such as "average," "inplace MC," "MC interp" [87]. To ease the explanation of these terms, assume two descriptions are created and streamed, but the decoder receives only one description, assume odd frames.

The "average" method is a simple algorithm to estimate even frames by making the average from its the adjacent odd frames.

The "inplace MC" method modifies the motion information of two consecutive frames of the available description (odd frames for this example). Then it scales motion vectors by a factor of $\frac{1}{2}$ and uses them as the motion vector for the unavailable description (even frames for this example).

The other method, called "MC interp," uses a phase correlation algorithm to

restore motion vectors of the missed frames [88].

The performance of the mentioned recovery methods depends on the correlation between the frames. For instance, a video with fast movement includes frames that are less correlated to each other and, therefore, missing one description causes an unacceptable side quality. To enhance the performance, Bai et al. introduced a scheme that adds frames during the fast movement period to increase the correlation between the frames [90]. Since the distance between available frames in the decoder is smaller than the original temporal MDC, it yields a better side quality; however, to display the real video, the mentioned extra frames are not displayed and are just used to hide the missed frames. The rate required to stream the video with extra frames (as presented in Bai's method) is greater than before. To decrease the rate, Zhang et al. have modified the method so that some frames during the slow periods of the video are discarded to keep the total rate approximately the same as it was in the original video [89]. The block diagram in Fig. 2.19 shows the structure used in [89].

Instead of a simple temporal assignment of frames between descriptions, there are more complicated methods, like the ones presented by Kibria et al. in [91] and Apostolopoulos in [87] that add motion vectors of the first description to the second description and vice versa. Both methods create only two descriptions, i.e., description one includes odd frames and description two includes even frames. According to results presented in [87, 91], if only one description is available at the receiver, the other description is estimated with better quality, as its motion vector information is available in that description.

Tillo and Olmo have also presented a similar temporal MDC algorithm in [92] to the spatial MDC algorithm that they had introduced in [73]. By this algorithm, both descriptions contain frames that are too hard to estimate from adjacent frames (i.e., frames including objects with fast movement). This way, the missed description can be estimated more accurately by simple linear interpolation.

Like spatial MDC methods, there are several multi-rate temporal MDC methods. For example, Tillo et al. presented a temporal multi-rate MDC with two descriptions. According to their work, each description contains a high frame rate of the first description and a low frame rate of the second description. If one description is dropped, a lower quality version of the missed frames is available as the side decoder. If both descriptions are received, the central decoder discards frames with a lower

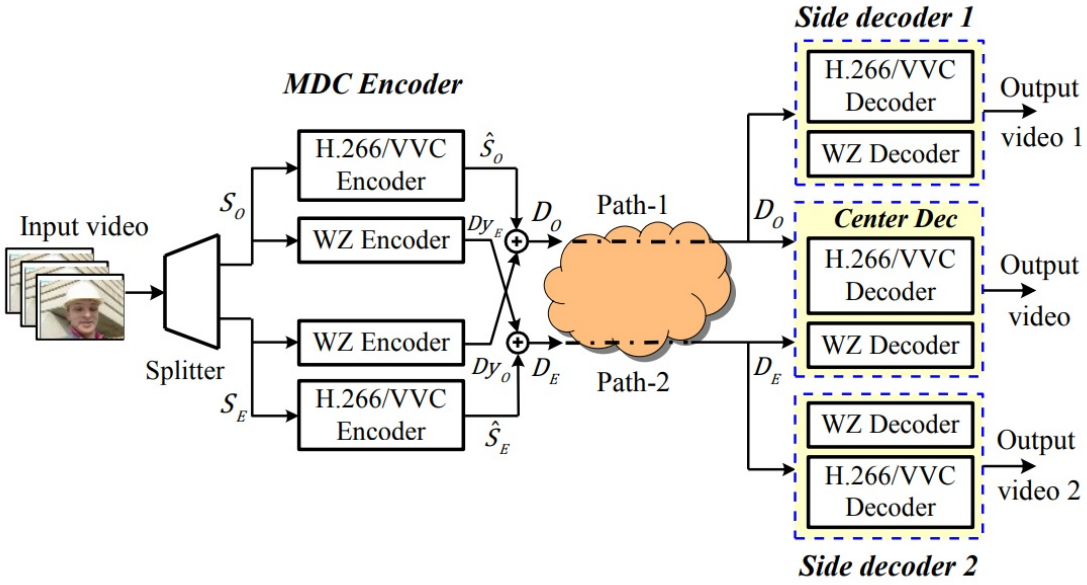
rate in a period equal to GOP length. Therefore, to switch between descriptions, the side decoder needs to wait for the end of GOP [93].

Another approach, described by Radulovic et al. [94], used an algorithm that is very similar to the one presented by Kibria et al. in [91]. The only difference is that each description is a lower rate of other descriptions (instead of including motion information of the other descriptions).

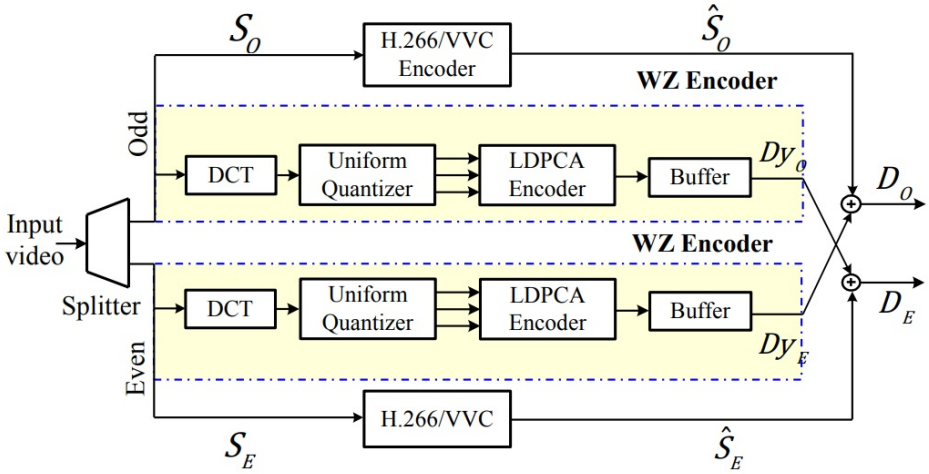
2.2.3.4.3 Frequency Multiple Description Coding

Compared to spatial and temporal MDC, the frequency multiple description coding approach divides the frequency components of a video among descriptions. There are several types of frequency MDC such as "multiple description scalar quantizer (MDSQ)" [95–97], "coefficient partitioning" [98–102], "multidescription transform coding" [103–106], and "frequency multirate MDC" described in the following.

In the multiple description scalar quantizer (MDSQ) method, each description needs to be quantized so that the combination of descriptions in the central decoder provides a lower quantization error. If we assume that there are only two descriptions, the quantization levels of the second description need to be shifted by half of the quantization step as seen in Fig. 2.20. Therefore the quantization step is smaller, and the quantization error is more significant in the central decoder than in the side decoder. H.264/AVC and H.265 [107–109] offer offset in the quantization levels via an adaptive quantization offset option.



(a) VVC-MDC method's block diagram [83] .



(b) Encode diagram [83].

Figure 2.17: VVC-MDC Encoder [83] ©IEEE 2020.

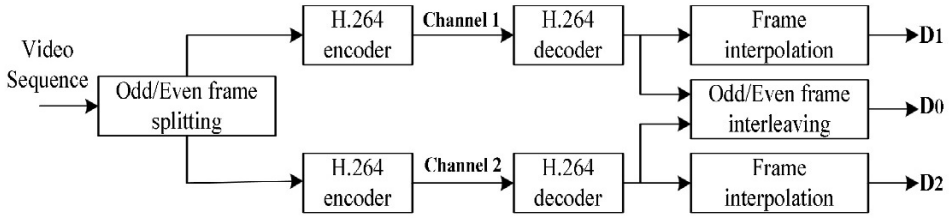


Figure 2.18: Temporal multiple description coding [89] ©IEEE 2008..

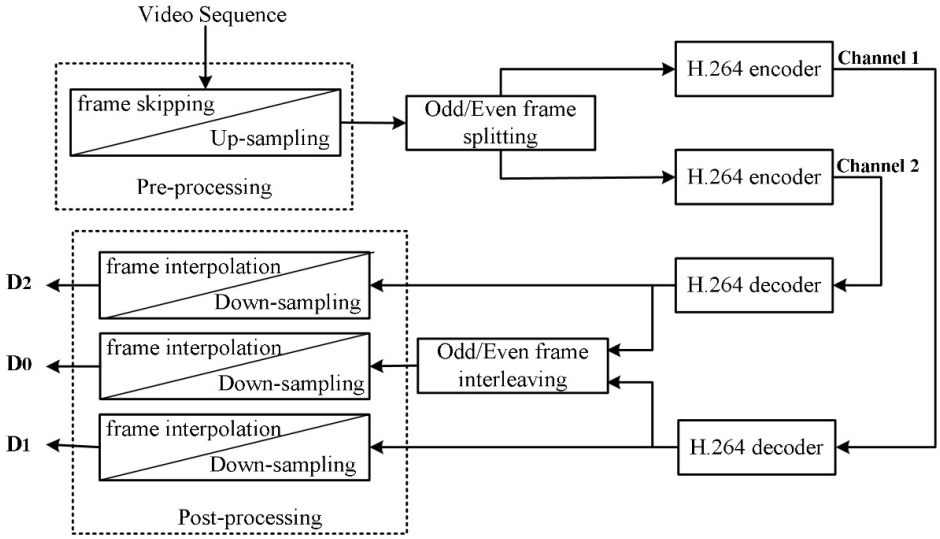


Figure 2.19: Temporal multiple description coding described by Zhang et al. [89] ©IEEE 2008.

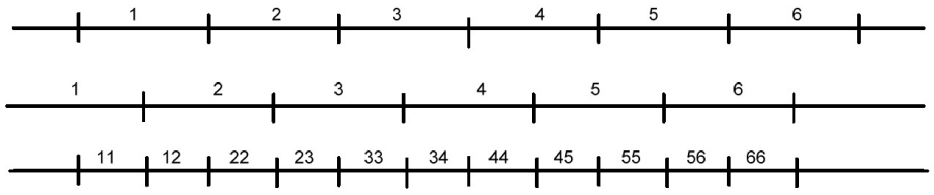


Figure 2.20: Quantization pace in side decoder (two above lines) and central decoder (last line) in MDSQ approach [95] ©IEEE 1993.

Parameswaran et al. modified this method so that not only the second description but also the first description have the quantization offset [95]. They also argued that the quantization offsets could be fixed or changed dynamically according to the type of video or application and the communication channel condition. Using an adaptive quantization offset provides better performance than a nonadaptive quantization offset; however, it is more complex.

One problem of the above MDC method is that the quantization offset needs to be less than half of the quantization step. Therefore, it creates a limit for the scenarios that only error-free or low error descriptions need to be received and displayed on the terminal devices. Such scenarios require that the encoder increases the redundancy to improve its resistance against the channel noise; however, it is not possible to increase redundancy without limit, as the quantization offsets are limited to the quantization step.

It should be noted that the quantization levels in the central decoder are not necessarily twice the quantization levels in the side decoders. Vaishampayan used this idea and introduced an algorithm such that the quantization levels in the central decoder are not as twice the quantization levels in the side decoders [96]. As shown in Fig. 2.21, there are eight indices for each description in the side decoders, and there are 21 quantization levels in the central decoder. Therefore, if unreliable communication happens and one description is missed, the side decoder has only the eight quantization levels. In contrast, if all descriptions are available, the central decoder has 21 quantization levels. Vaishampayan completed his algorithm later in [97] and found the optimum indices allocation table for the input with uniform distribution. Tian and Hemami extended Vaishampayan's algorithm and introduced a two-step quantization scheme compared to Vaishampayan's work [97]. According to Tian and Hemami's method, it is better to have an extra step to quantize the quantization result of the first step, again. In this scheme, all descriptions have quantization levels equal to the quantization levels in the first step, but the second quantization indices are distributed between descriptions. Therefore, as the side decoders do not have access to the quantization indices assigned in the second step completely, the side decoders ignore them and only consider the quantization levels assigned in the first step. In contrast, the central decoder has access to all quantization indices assigned in the second step and therefore provides a minor quantization error.

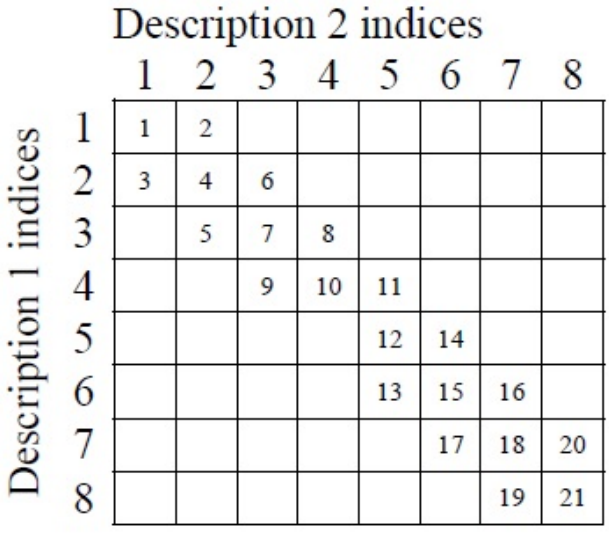


Figure 2.21: Modified index assignment described by Vaishampayan for MDSQ multiple description coding [96] ©IEEE 2004.

As can be understood from the term "coefficient partitioning," this method divides the DCT coefficients among descriptions. There are several algorithms that use this method presented in [98–102]. According to Reibman et al. [98], first, a threshold for DCT coefficients is defined and then the coefficients that are greater than the threshold are assigned to all descriptions. Other coefficients, which are smaller than the threshold, distribute among descriptions to enhance descriptions' quality. Therefore, a description includes all large DCT coefficients and some small DCT coefficients. Each description includes a part of the small DCT coefficients, and the other small DCT coefficients are replaced with zero in that description. Therefore fewer bits are needed for the encoding due to the "run of zero" algorithm, although distortion in the side decoder is greater. As the central decoder has access to all DCT coefficients, distortion is less and it provides better quality. Matty and Kandi argued that the threshold is the adjustment tool to control the redundancy-distortion problem [99]. They found the threshold's optimum value to balance the rate-distortion trade-off problem [99].

Comas et al. introduced another coefficient partitioning method, creating descriptions non-symmetrically [100]. Based on their method, only one description, called the main description, includes all the coefficients, and all other descriptions contain only the critical coefficients [100]. This procedure can save more bits due to the

"run of zero" algorithm; however, its robustness against noise is not as good as other symmetrical MDC methods, as it is likely to miss the main description.

DCT coefficients can also be partitioned into several descriptions in a block-wise manner; however, it provides poor side quality since the estimation of a block from another block is not accurate. To avoid such degradation in the side quality, "lapped orthogonal transformation" needs to be applied on adjacent blocks to increase the correlation between blocks. Chung and Wang used the lapped orthogonal transformation after the discrete cosine transformation (called DCT-LOT transformation) and then distributed the transformed blocks between the descriptions [101]. Guoqian et al. used a similar idea [102] to partition the frequency coefficients into the descriptions. The only difference that they made compared to Chung and Wang's method is adding more redundancy to increase the side quality. The extra redundancy, which is only added to one description, comes from the difference between the blocks assigned to the different descriptions.

DCT coefficients of each block cannot estimate accurately from other coefficients of the same block, as coefficients are orthogonal. In other words, DCT transformation transfers the block's pixels from the spatial domain into the frequency domain in which its components are independent of each other. Indeed, video coding standards use transform coding such as DCT or wavelet transformation to eliminate the common information between blocks' pixels and consequently increase the compression efficiency. Therefore, if the DCT coefficients of a block are distributed between various descriptions, the DCT coefficients of a block in the missed description cannot be estimated from the DCT coefficients of the same block of other descriptions. One solution to avoid this problem is to estimate the frequency coefficients of the missed description from the corresponding frequency component of the adjacent block available in other descriptions. Although this method outperforms the method that estimates the missed frequency coefficients from the other frequency coefficients of the same block, the quality of the reconstructed video in the side decoder is not usually high enough, and its performance depends highly on objects' movement. The other solution to conquer the independence of frequency coefficients of a block, as suggested by Wang et al. [110, 111] and continued by Goyal et. al in [103, 112] is using a correlating transformation to increase the correlation between the block's frequency components. This procedure, known as "multi-description transform coding" (MDTC), applies another transformation, called pair-wise correlating transformation

(PCT), just after the DCT transformation. Then, the transformed coefficients are assigned to the different descriptions. Indeed, pair-wise correlating transformation changes two uncorrelated coefficients to two correlated coefficients. This helps the estimation algorithm in the side decoder find the lost coefficients from other available coefficients. Like other MDC algorithms, this one also enhances the side quality at the expense of the coding compression ratio. It should also be noted that the complexity is increased with MDTC as an extra transformation, i.e., PCT, is required.

To moderate the decreasing rate of the compression ratio that happens with the multi-description transform coding algorithm, one solution, as suggested by Goyal et al. in [103, 112] and Wang et al. in [104], is to apply the pairwise correlating transformation to a limited number of coefficients, not all coefficients. In this case, the pairwise correlating transformation needs to be applied only to those coefficients that do not have the same distribution, as argued in [103, 104]. Wang et al. have also shown that this approach is not valid for the low error applications, which need less redundancy [105]. Wang et al. improved this method's performance for the low error applications by updating the previous algorithms presented in [104, 105] and adding one more step that is copying the least predictable parts of coefficients into all descriptions [106, 113].

Sampling-based multiple description coding is more flexible compare to transform-based multiple description coding, however they are manually designed for a specific sampling. Lijun et al introduced a deep multiple description coding framework using convolutional neural network to leverage images context features when creating descriptions [114]. Later they modified their work to improve the coding efficiency as it was not suited for the practical applications [115]. Due to high complexity of their method and also learning, this type of multiple description coding is not beneficial for live streaming video specially for long distance networks.

Similar to spatial and temporal multi-rate MDC methods, the multi-rate frequency MDC method is also feasible. The multi-rate frequency MDC can be applied to DCT coefficients either element by element or block by block. For example, the DCT coefficients can be quantized with different quantization paces to produce video streams with different rates. Samarawickrama et al. proposed a multi-rate frequency MDC algorithm so that each description contains half of the high-rate-quantized coefficients and half of the low-rate-quantized coefficients (they have assumed that there are two descriptions) [116].

Compared to the method presented in [113], Wang et al. introduced a multi-rate frequency MDC algorithm in which different quantization paces are applied to the DCT coefficient blocks [117]. With this algorithm, every other DCT coefficient block in each description is quantized with the same quantization pace. For example, in the first description, the odd blocks are quantized with the small quantization pace and others are quantized with the large quantization pace. The second description contains the odd blocks with the large quantization pace and the even blocks with the large quantization pace. Therefore, both descriptions contain all DCT coefficients of a frame. To improve the compression efficiency, Su et al. argued that motion vectors of the low rate block are not required to be streamed, and they can be estimated from the motion vectors related to the high rate blocks [118]. To ease the implementation of this algorithm, Tillo et al. used the redundant slice option available in H.264 to encode blocks with a large quantization pace [119]. It is worth mentioning that the redundant slice is a feature available in the H.264 standard to encode each slice with the lower rate and replace it with the original slice in the case of missing a slice.

Tian and Rajan have suggested using various transformations for each description [120]. For example, the DCT transformation is applied for only one description, while no transformation is used for the other description. It should be noted that such a pair of transformations is just a sample and clearly, there is a better combination of transformations to improve the result. It is worth mentioning that the descriptions created with this method are not symmetric and the rate and error in one side decoder are not the same as the other one.

2.2.3.4.4 Hybrid Multiple Description Coding

However, multiple description coding methods provide a reliable video stream; it also increases the redundancies and decreases the compression ratio. As explained before, such a decrease in the compression ratio is caused due to the:

- extra header, as every description is required to have it separately.
- decrease in the dependency of data in each description and consequent decrease in the DPCM algorithm's performance because similar data is not assigned to one description.

So, extending the number of descriptions to more than two decreases the compression ratio dramatically. Therefore, MDC methods usually create two descriptions [63, 70].

For the scenarios in which more descriptions is required, one solution is to select an MDC algorithm that has less effect on the compression ratio, although it is not very reliable against the packet loss.

The other solution is to use a hybrid MDC method. This way, the decreasing rate of dependency between data in one description is mitigated, as each type of MDC targets a different domain. For example, Lu et al. introduced a hybrid MDC algorithm that combines temporal and spatial domains to create descriptions [121]. Based on Hsiao and Tsai's MDC algorithm, the spatial MDC and frequency MDC (coefficient partitioning) are combined [122]; Xu et al. also have combined the spatial MDC and frequency MDC. The only difference is that they chose the multiple description scalar quantizer in order to use the frequency domain [123].

Yu and Jin analyzed advantages and disadvantages of both temporal and spatial multiple description coding and provided an efficient hybrid MDC algorithm using the H.264 video coding standard [124]. They compared the result of different multiple description coding methods with four descriptions, including temporal MDC, spatial MDC, and hybrid spatial-temporal MDC and showed that better performance can be achieved by the hybrid MDC algorithm. According to this work, the temporal MDC method provides better performance against network failure with a specific power for the noise compared to the spatial MDC method; however, the temporal MDC method is more sensitive to the variance of the noise. As Yu and Jin's simulation result shows, the hybrid MDC method can have a performance as good as the performance of the temporal MDC method while the performance decreases gradually with the increment of the noise power just like the spatial MDC method.

2.2.3.4.5 3D/multiview Multiple description coding

Similar to 2D MDC, 3D/multiview multiple description coding aims to avoid packet loss of noisy communication networks. The only difference is different views can be assigned to the separate descriptions. For example, a simple 3D MDC method can allocate left view of an stereoscopic video to one description and assign the right view to the other description.

Jing et al. introduced a new multiple description coding for multi view video in which description are created for each view separately. Data of each description is assigned with an adaptive redundancy allocation algorithm [125]. As can be seen in Figure 2.22, the multi-view video is downsampled by polyphase transform in spatial

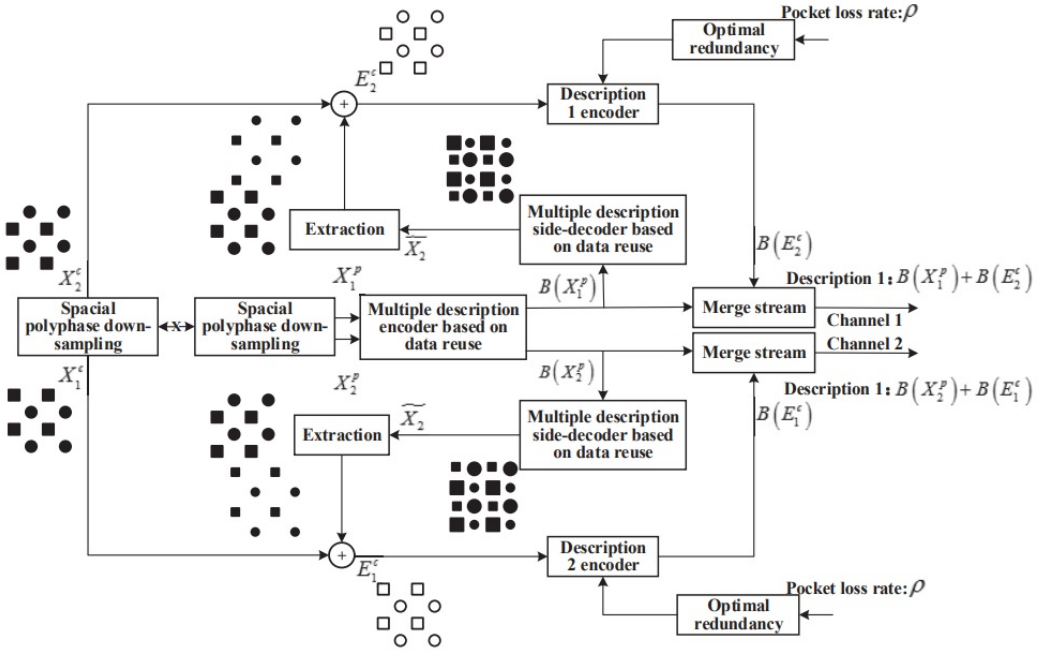


Figure 2.22: Multiview multiple description coding with adaptive redundancy allocation [125] ©IEEE 2017.

domain. The difference between the side decoder and central decoder are encoded to form the redundant part and transmit with the original description. In receiver, if both descriptions are available central decoder is chosen to decode otherwise central decoder is used to reconstruct video using the received base and redundant parts. Again, number of descriptions for each method is limited to two descriptions for each view. Also descriptions are made separately for each view and the decoder does not exploit of the correlation between adjacent views.

Norkin et al. introduced an MDC method applicable on stereoscopic videos [126]. It creates descriptions through temporal subsampling (see Fig. 2.23). In this algorithm, called a multistate Stereo-MDC (MS-MDC) Scheme, each description contains even or odd frames from both views. Also, a motion compensation process in one view predicts from the second previous frame and the other view, while the other view is predicted only from the second last frame. As shown by Norkin et al. in [126], spatial scaling stereo-MDC (described previously) outperforms in scenarios in which there is a low correlation between the left and right views. For the scenarios where there is a

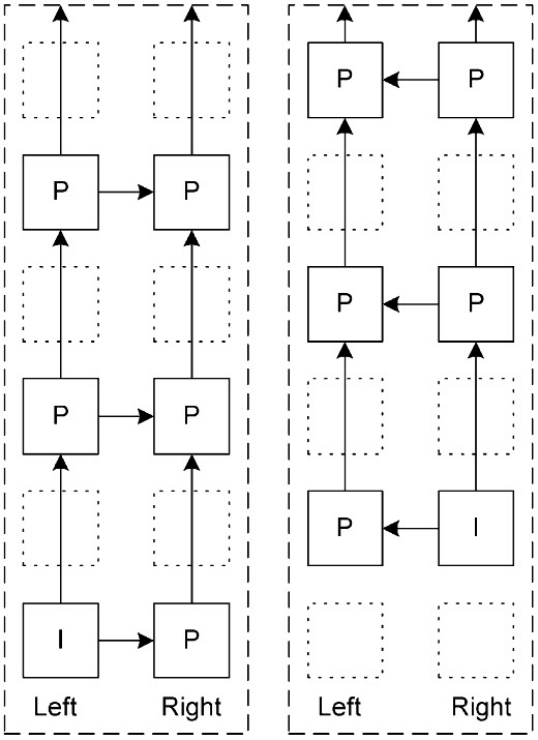


Figure 2.23: Descriptions of multistate Stereo-MDC Scheme. Reprinted by permission from Springer: Springer Multimedia Content Representation, Classification and Security [126] ©2006.

high correlation between the left and right views, the simulation result in [126] shows that the multistate Stereo-MDC Scheme provides better performance.

It is worth mentioning that polyphase subsampling has excellent performance for the depth map image, as it includes very correlated and also low-frequency data. Karim et al. in [127] used such characteristic of the depth map image and introduced a multiple description coding algorithm with a better performance at the point of compression ratio for the stereoscopic video. They compared the reconstructed 3D video's quality using the original depth map image and the downsampled version of the depth map image. They concluded that the decimation of the depth map image does not cause any considerable degradation in the decoded 3D video.

2.2.3.4.6 Scalable Multiple Description Coding

As explained in the previous and current chapters, the scalable encoding and multiple description coding methods address different issues of the video streaming; therefore,

a scalable multiple description coding method can gain both methods' advantages and produces a scalable video stream that resists packet failure. For example, a common problem that often happens with the spatial multiple description coding is sharp edges in the scene or scenes with thin and small objects. Since spatial MDC partitions frames spatially, such small objects might disappear in one or some descriptions. The scalable multiple description coding algorithm described by Choupani et al. addresses this problem [128]. As the simulation results show, this problem is not sensible for the multiple description coding method with two descriptions; but for the multiple description coding method with more than two descriptions, their algorithm provides an acceptable performance. They showed that their algorithm for the multiple description situation with more than two descriptions in a powerful noise-prone environment provides about 1 dB better performance in the rate of 1 Mb.

Basically, the scalable MDC approach can be applied to video via two distinct algorithms [127]:

- To produce a scalable MDC stream, first a scalability algorithm can be applied to the video stream and then layers can be partitioned into several descriptions [127,129–132]. As argued before, the lower layers in the scalable video encoding are more important than the higher layers. This is due to the dependency of higher layers on lower layers for the decoding process. This means, in the case of packet loss in the last layer, only bits related to that layer will be wasted; while receiving errors in the base layer means that the upper layer also cannot be decoded and bits for all layers will be discarded. As the lower layers are more important than the upper layers in the scalable coding, the multiple description coding algorithm can be applied only on the lower layers wisely. For example, the scalable multiple description coding algorithm proposed by Karim et al. creates a scalable stream at first, i.e., base layer and enhancement layers and then separates even and odd frames of the base layer to apply a multiple description algorithm [127]. Therefore, they made the base layer error resilient by applying the MDC algorithm on the base layer.

- Another algorithm to produce a scalable MDC stream is applying MDC and scalability in the reverse order. In other words, first, a video's frames can be divided between descriptions and then a scalable algorithm needs to be applied on each description [133–135]. For example, in the hybrid scalable multiple description coding algorithm introduced by Favalli and Folli in [135],

the scalable encoding algorithm has been applied to the descriptions obtained from the multiple description coding. Indeed, four subimages are generated via spatial subsampling and then each of two subimages are combined by a scalable algorithm to create one description. Therefore each description contains one base layer and one enhancement layer and each sub-image is assigned to each layer (base or enhancement layer); this way a video stream can be transmitted by a multiple description coding method and in the case of network fluctuation, the resolution of each description can be decreased to adapt to the network changes. Also, as sub-images are very similar to each other, using the interlayer prediction that is available in the SVC encoder assures the compression efficiency.

2.2.3.4.7 Performance Assessment of Multiple Description Coding Methods

Having examined a wide variety of multiple description coding algorithms, the advantages and disadvantages of different multiple description coding methods must be compared to find the most appropriate method for an application. To this end, first one needs to see what aspects of multiple description coding algorithms must be assessed for a special application. These aspects are as follows:

- **Capability of controlling on redundancy:** As described, an application may be used by a user located in a low noise condition. Therefore, the descriptions are often delivered error free and there is no need to have a large amount of redundancy bits for each description. On the other hand, in some scenarios, the application may be used in an error-prone environment. Therefore, using more redundancy bits to secure the descriptions is sensible. So, a multiple description coding needs to have the ability to increase or decrease redundancy to stream a video more effectively. For example, as mentioned before, the polyphase subsampling MDC algorithm has a poor ability to control the redundancy precisely, and the zero padding MDC algorithm introduced by Shirani et al. fixes this problem of polyphase subsampling MDC algorithm [71, 72].

- **Possibility of increasing the number of descriptions:** As argued earlier, multiple description coding algorithms usually create only two descriptions because having more than two descriptions drops the compression ratio; however, in some scenarios in which the noise variance is too high, more descriptions are required. Additionally, for some scenarios such as multipath transmission [136],

point-to-point video transmission more numbers of descriptions are required as a variety of rates and resolutions are required for the different paths or different receivers, respectively. Generally, the possibility of increasing the number of descriptions for the frequency MDC method is higher than the temporal and spatial MDC methods because with increasing the number of descriptions, the dependency of data in spatial or temporal domain drops significantly.

- **Complexity:** Complexity is a very important concern for video streaming, especially for applications dealing with instantaneous streaming, such as video conferences. While processing power has increased in the last few years (according to the Moor's law, processing speed doubles approximately every two years [137]), new video coding standards also need more processing power. For example, according to [138], H.264 needs more processing power in comparison to its former video coding standards. About HEVC, it can be said generally decoding software has been simplified while the encoder side needs powerful processing power to make a real time encoding [139]. In addition to the processing power required for encoding and decoding, multiple description coding adds more processing load to create the descriptions. For example, the multiple description coding algorithm introduced by Yapcthe et al. in [74] is not suited for live streaming, as it needs more processing power to design the filter coefficients.

- **Sensitivity to the packet failure or noise:** As network conditions fluctuate continuously, no streaming method can always guarantee to deliver the packets without error, although the fluctuations in some networks are low and in some others very high, depending on the environmental condition, number of users in the network, location of user, etc. Therefore, a proper multiple description coding algorithm should not be affected by the increase in the variance of the noise. For example, as described before, temporal MDC methods are more sensitive to the increase in noise power than spatial MDC methods [124].

2.3 State of The Art

As explained in the previous chapter, with the MDC method a video data stream is partitioned into several descriptions and then encoded separately. The descriptions

are then streamed through the network toward receiver(s). In the receiver, there are two different types of decoder - the side decoder and central decoder. The receiver chooses one of the two decoders based on the availability of the error free descriptions. The instance that all of the descriptions are received successfully is when the central decoder is activated. Otherwise, the side decoder will be activated when only a few error free descriptions are received.

The MDC method is best recognized for its robust error behavior at the expense of compression ratio as it adds redundancies in its temporal, spatial or frequency domain. With the temporal MDC method, only two descriptions are usually created to avoid a huge drop in the coding efficiency. The drop in the coding efficiency is reflected more (when greater than two descriptions are used) because the distance between the assigned frames to each description is increasing resulting in the motion prediction being less effective [63, 70]. The higher noise of the network that video are streamed through, MDC methods with more number of descriptions outperforms. Therefore the temporal MDC method is no longer a suitable technique. The frequency MDC method partitions Discrete Cosine Transform (DCT) coefficients between video descriptions. Because DCT transformation provides independent components, the descriptions will be less dependent. To maintain the correlation of the descriptions, an extra transformation like Lapped Orthogonal Transformation (LOT) needs to be applied. Therefore the complexity of frequency MDC methods is higher than that of both the spatial and temporal MDC methods. With the spatial MDC method, each video frame is partitioned into several lower resolution subimages using Polyphase SubSampling (PSS) algorithm [27, 71, 72]. It is worth mentioning that with a simple spatial MDC method, there is no precise adjustment tool over the redundancy in order to control the side quality [27, 71, 72]. This means that there is no control for the redundancy increase resulting in higher resistivity to compensate for the higher noise level.

To improve the basic spatial MDC method's, Tillo and Olmo introduced a new MDC algorithm called "least predictable vector directional multiple descriptions coding" [73]. This approach basically copies the least predictable parts of the frame to all descriptions. Their simulation result shows that this method improves the side quality when compared to previous specially down-sampled MDC method although the new method provides more redundancy. Tillo and Olmo obtained better quality for higher noise level at the expense of less coding efficiency and greater algorithmic

complexity.

Shirani also presented a nonlinear PSS-MDC method which investigated its performance by evaluating the case where there were one or more missing descriptions [76]. According to his work, some parts of a frame which are more important called Region Of Interest (ROI) were sampled with a greater rate (based on an exponential equation) compared to other parts of the frame. In other words, descriptions include more information regarding the ROI parts of the frame resulting in an enhancement of the side quality. More importantly, this method provides for greater performance with regards to the subjective assessment by the human eye since objects and not pixels are more emphasized. Although Shirani's method provides for the enhancement of the side quality, he did not discuss how the ROI parts of a frame were detected which is important when involving fast video contents or live video streaming. In this paper, we provide a new spatial MDC algorithm that adds redundancy to the descriptions more practically for 3D videos.

To apply the MDC method for 3D videos, the depth map image also needs to be partitioned into different descriptions. It is worth mentioning that the depth map image mainly contains depth information of the scene's objects. Because of the nature of the real objects, depth information of 3D scenes rarely contains high-frequency content. Consequently, the depth map image can be effectively compressed resulting in saved bandwidth and disk space [2,50]. To improve compression, Karim et al. have shown that the downsampled version of the depth map image provides an adequate reconstruction of the 3D video in the receiver [127]. They have experimented with the spatial MDC method for 3D videos using color plus depth map image representation. Karim et al. have carried out experimental tests with a scalable multiple description coding approach arriving at the same result. Therefore, it can be said that downsampling of the depth map image does not cause a considerable degradation in the quality of a reconstructed video. This is due to the fact that the depth map image includes low-frequency contents or more precisely, the depth values of adjacent pixels are similar. Consequently, one can state that the neglected pixels during downsampling can be better predicted. Liu et al utilized the fact of having similar depth values of pixels for real objects and introduced a texture block partitioning algorithm in order to perform their MDC algorithm for wireless multi-path streaming [70].

Chapter 3

Reliable 3D Video Streaming Considering Region of Interest

This section describes the new multiple description coding applicable to 3D videos and was already published in [140, 141].

As argued in the previous chapter, the most common multiple description coding type is the one that allocates frames to descriptions temporally, because it has low complexity and it is very simple. However, the number of descriptions cannot be increased due to compression inefficiency. Additionally, for live streaming, waiting for the future frames to be able to decode the current frame is not the best choice when the transmission delay is impotent and also we need more than two descriptions. In this thesis we are looking to implement our MDC method in the spatial domain as this type doesnot need to move forward and backward between the frames and less delay is required to decode the video. However, the receiver reconstructs video at a lower quality compared to the temporal MDC. To improve the performance of spatial MDC, we are creating descriptions spatially while assigning more bandwidth to region of interest (RoI), or important objects of the scene.

To realize the RoI of a frame we need to define a metric to identify RoI bloks and clarify how it can be used for the purpose of RoI extraction. Briefly, the new MDC method consists of three steps:

- Map extracting for RoI: First, each 3D raw frame is split into a 2D color frame and a gray scale depth map frame. Then we are looking for different regions of the frame using the depth map image. The process of extracting the RoI is described in Section 3.2.1.1. It is worth mentioning that is also possible to used other saliency methods to detect RoI. In this case, the RoI definition may be

different from what is defined by this thesis.

- Creating description using polyphase subsampling: four subimages are created using polyphase subsampling (PSS) from both the color and the depth frame separately as explained in Section 3.2.1.2,

- Enhancing the descriptions: Enhancement of the descriptions is achieved through the combination of different regions of the frame with different resolutions obtained from the color and depth map streams. This step of the new MDC algorithm is fully described in Section 3.2.1.3 and Section 3.2.2.1.

3.1 RoI Extraction Metrics

We define the RoI of a frame as those objects of the frame that are usually in sharp focus during recording. In photography, sharp photos means that you want an object to be in focus with clear lines and no blurring. In other words interesting objects are those that whose depth is not very far from the camera compared to other parts of the frame. It is worth mentioning that when a video is recorded, important objects are usually selected to be in sharp focus, and the background is slightly blurred. In general, a photographer changes depth of field to control how much of a frame is in sharp focus. After, extracting the map for interesting objects, we divide a frame into three regions and change the bandwidth for different regions. We call these regions the following:

- region I, or, the background,

- region II, or, the interesting objects,

- region III, or, the edge of those interesting objects.

To identify the RoI of the scene, we evaluate two metrics from the second-order statistics of the depth map image in a block-wise manner, and compare their performance.

Generally, the depth information of an object includes low-frequency content data as an object can only be located in one place at a time. Therefore, the RoI can be extracted by filtering low-frequency parts of the frame which aren't located very far from the camera. To this end, we will introduce a hierarchical block division (HBD) algorithm as presented in [140] to find areas with similar depth values (which are

indicated as low-frequency data). Metrics used in this thesis to extract RoI map are *Coefficient of Variation(CV)* and *Variance (PV)*. We will demonstrate that CV outperforms PV as shown by the simulation results presented in the next chapter. It is worth mentioning that CV is defined as the ratio of the standard deviation to the mean [142, 143], and calculated for each block as shown in Equation (3.1).

$$CV^{B_k^l} = \frac{\sigma^{B_k^l}}{\mu^{B_k^l}}, \tag{3.1}$$

where B_k^l stands for k^{th} block in l^{th} iteration of HBD algorithm (HBD algorithm will be explained in the next section); k varies from one to the total number of blocks in each iteration. For example in the first iteration, the total number of blocks is one. We will talk more about the range of l and k later. $\sigma^{B_k^l}$ and $\mu^{B_k^l}$ are also the standard deviation and the average of k^{th} block in l^{th} iteration of the depth map image and are shown in Equation (3.2) and Equation (3.3), respectively.

$$\sigma^{B_k^l} = \sqrt{\frac{1}{N^{B_k^l} M^{B_k^l}} \sum_{i=1}^{N^{B_k^l}} \sum_{j=1}^{M^{B_k^l}} (d_{ij}^{B_k^l} - \mu^{B_k^l})^2}, \tag{3.2}$$

and

$$\mu^{B_k^l} = \frac{1}{N^{B_k^l} M^{B_k^l}} \sum_{i=1}^{N^{B_k^l}} \sum_{j=1}^{M^{B_k^l}} d_{ij}^{B_k^l}, \tag{3.3}$$

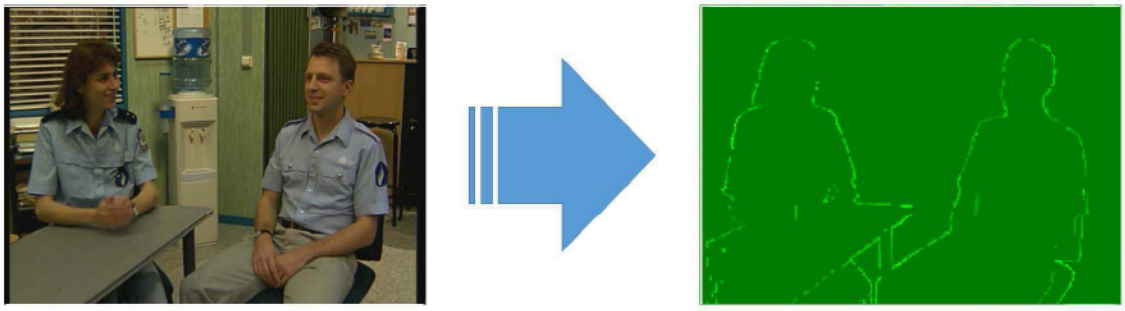
where $N^{B_k^l}$ and $M^{B_k^l}$ are the number of columns and rows of k^{th} block in l^{th} iteration (B_k^l), respectively. $d_{ij}^{B_k^l}$ is the depth value of the pixel located at column i and row j of block B_k^l.

The alternative metric used in this thesis is pixel variation or dispersion of pixels' depth values in the block B_k^l, called PV. PV of block B_k^l is calculated as Equation (3.4).

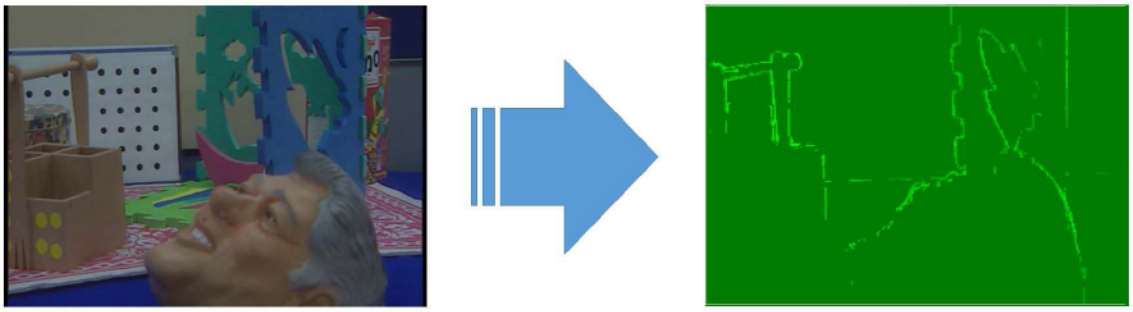
$$PV^{B_k^l} = \frac{1}{N^{B_k^l} M^{B_k^l}} \sum_{i=1}^{N^{B_k^l}} \sum_{j=1}^{M^{B_k^l}} (d_{ij}^{B_k^l} - \mu^{B_k^l})^2, \tag{3.4}$$

where $\mu^{B_k^l}$ is the average of k^{th} block in l^{th} iteration of the depth map image, and is already shown in Equation (3.3).

Generally, PV of a block is a positive value and indicates how much depth value of pixels of a block are close to the mean or spread out over a wider range. Blocks with large PV are probably related to several objects or edges. We label this part of the frame as region III. In Figure 3.1 pixels related to region III of the first frame of both the videos "Interview" and "Orbi" are shown.



(a) Video "Interview"

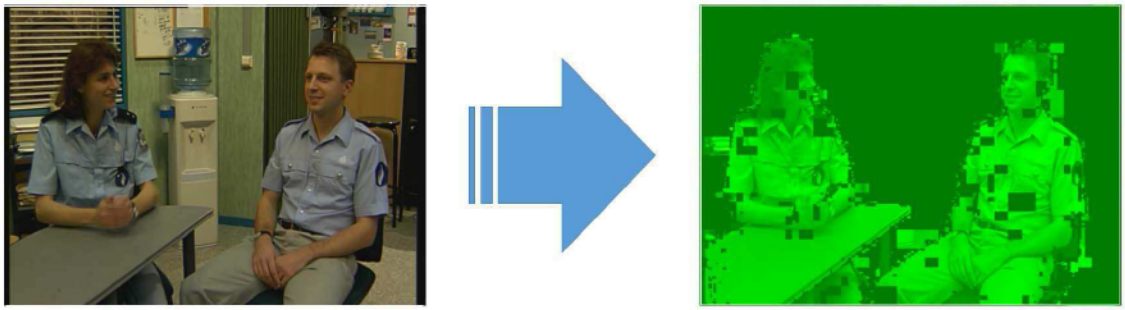


(b) Video "Orbi"

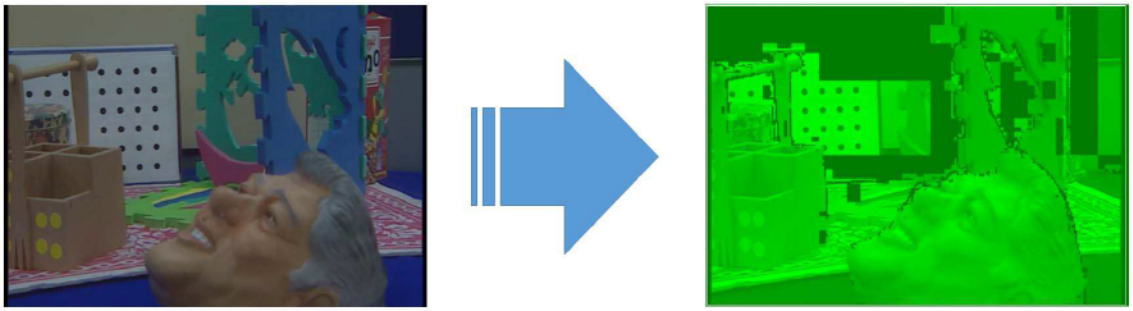
Figure 3.1: Example of region III .

Because the depth information of an object contains low-frequency contents naturally, the pixel depth values of an object are similar. This part of the frame is called Region II. Region II in Figure 3.2 is showing the objects of interest depicted from the videos "Interview" and "Orbi", respectively.

Blocks with very small PV (explained later) are related to the far distanced background or the planar objects, for example, a wall. We call this part of the frame as region I. Figure 3.3 shows the area related to the Region I of the first frame of both videos, "Interview" and "Orbi".

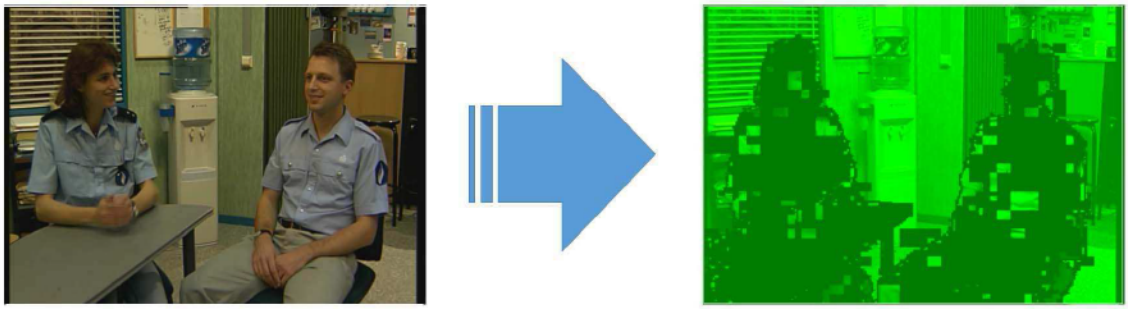


(a) Video "Interview"

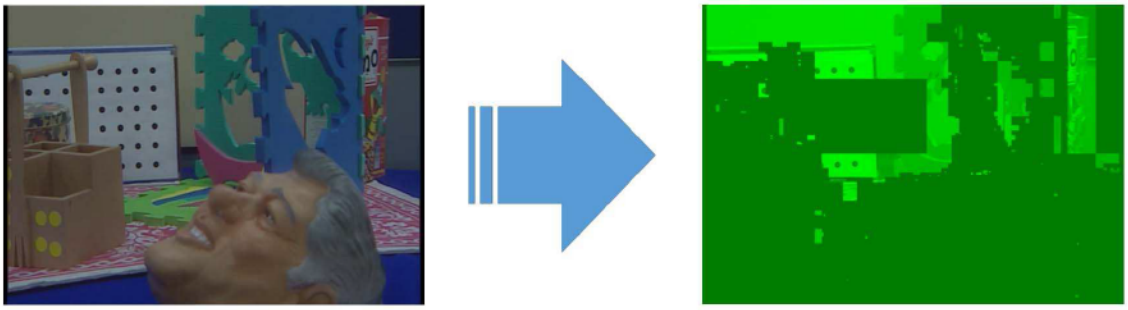


(b) Video "Orbi"

Figure 3.2: Example of region II .



(a) Video "Interview"

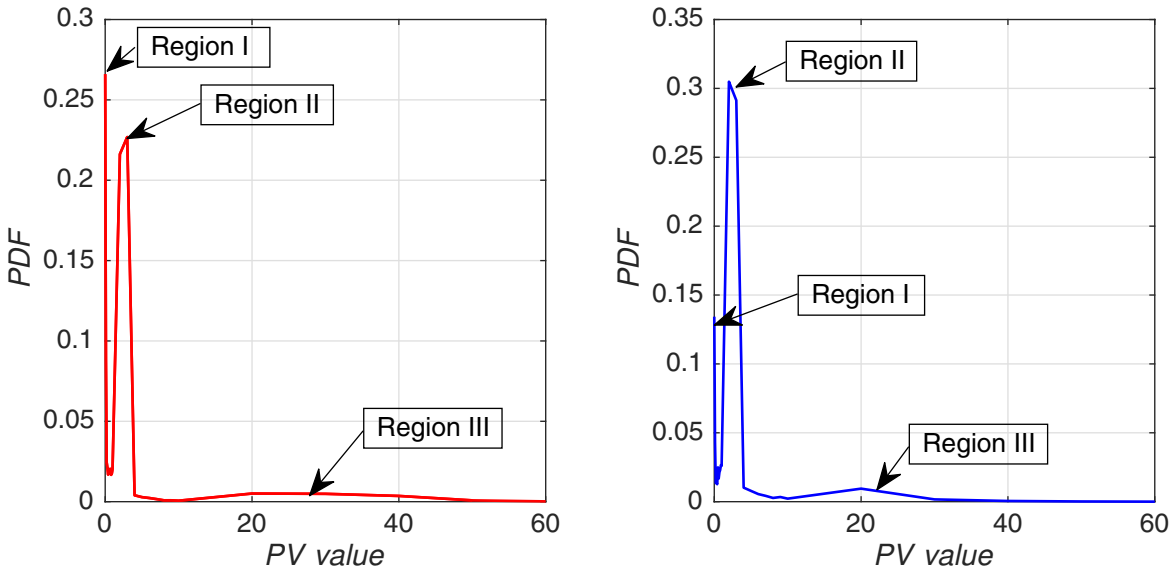


(b) Video "Orbi"

Figure 3.3: Example of region I .

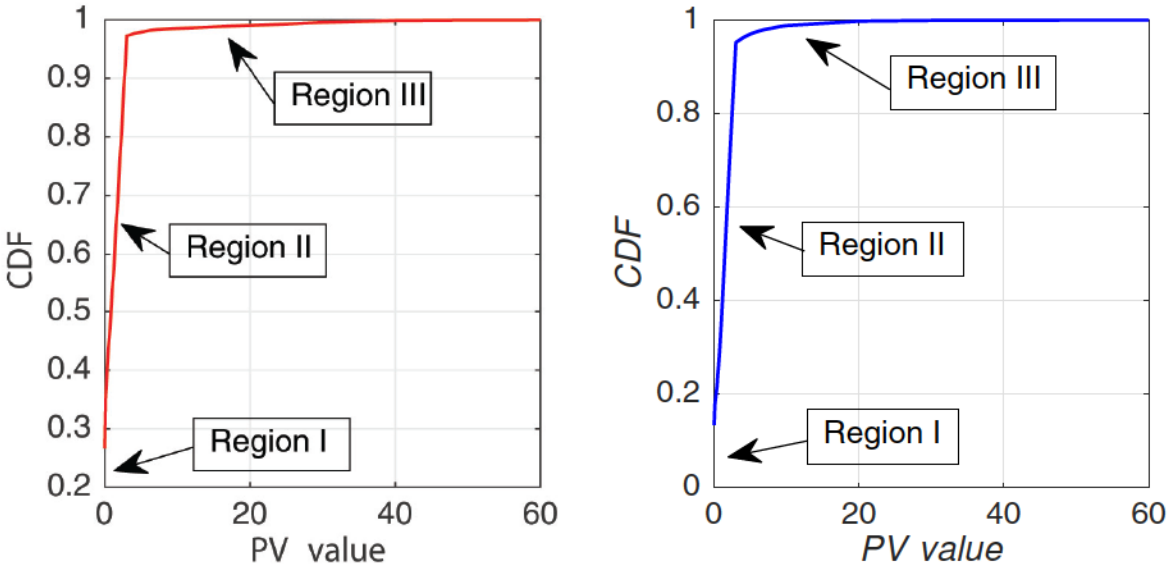
Figure 3.4 and Figure 3.5 show the probability density function (PDF) and the cumulative density function (CDF) of PV for two sample videos entitled "Interview" and "Orbi". In Figure 3.4, the PDF ($Pr(PV = x), 0 < x < +\infty$) of the PV values for these two videos depict pixels that are classified into the following three regions: regions I, II, and III. region I demonstrates that pixels of the frames have very low depth variation. Due to this very low variation region, the CDF is shown to start from a nonzero point in Figure 3.5.

Pixels related to region III are shown in Figure 3.4 as a small peak when the PV value is greater than 5. In Figure 3.5, region III starts from the point that the slope of the CDF graph changes from steep to moderate. region II in Figure 3.4 is related to the second peak and CDF in Figure 3.5 for its steep slope. For the remainder of this thesis, regions I, II, and III will be named background region, region of interest (or interesting objects' region) and edges region, respectively.



(a) Video "Interview" **(b)** Video "Orbi"

Figure 3.4: The probability density function of PV for the first frame of video "Interview" (left) and "Orbi" (right). As can be seen, three regions are distinguished for both video tests. These regions show blocks with depth variation very closed to zero, between zero and one, and greater than one (approximately)



(a) Video "Interview" **(b)** Video "Orbi"

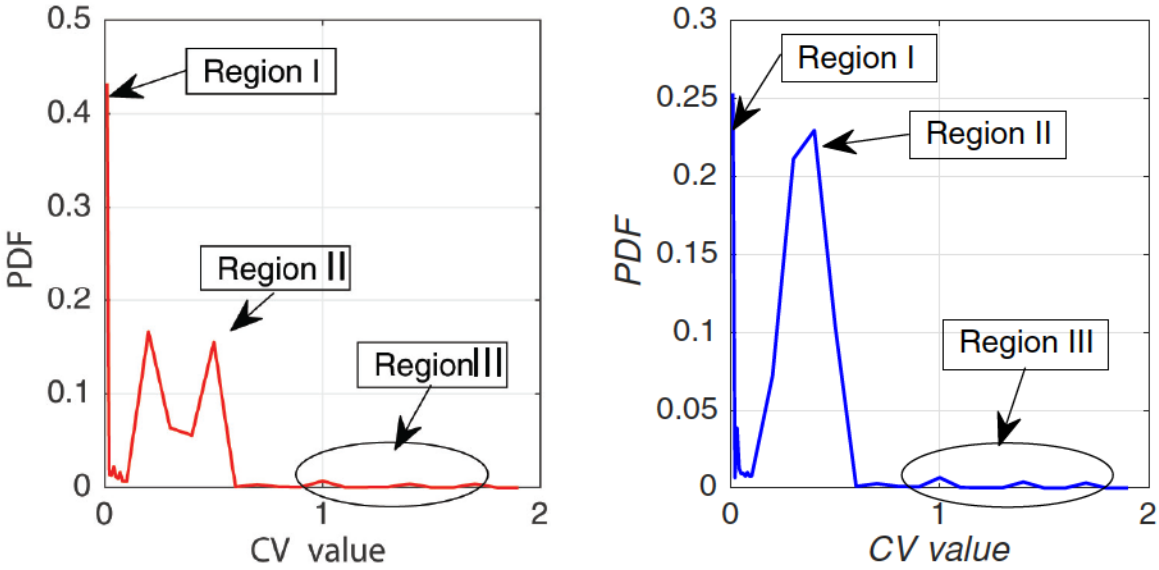
Figure 3.5: The cumulative density function of PV for the first frame of video "Interview" (left) and "Orbi" (right). Like its PDF, three different regions can be recognized for both video tests.

In comparison to PV, the CV of a block also varies from zero to infinity. As discussed in [144] and [145], CV can be used as a measurement tool of the level of heterogeneity or clustering for a random process. They showed that a realization of a process with a *super-Poisson* characteristic results in a CV greater than one, while a *sub-Poisson* characteristic results in CV less than one, and a *Poisson* characteristic results in CV equal one. Because the depth values of all pixels for an object are very similar, the CV metric can be used as a benchmark metric to determine the level of clustering and consequently able to extract objects from the depth map image.

As can be seen in Figures 3.1 to 3.3, the performance of the RoI extraction algorithm is not high enough. For example in Figure 3.2, there are some missing blocks in the middle of interesting objects; or some blocks have been detected as RoI in the background which are in fact not interesting objects. This is due to the fact that it is not accurate to use standard deviation of different blocks (see Equation (3.2)) for evaluation and comparison. In other words, pixel variations of respective blocks found in different scales need to be normalized.

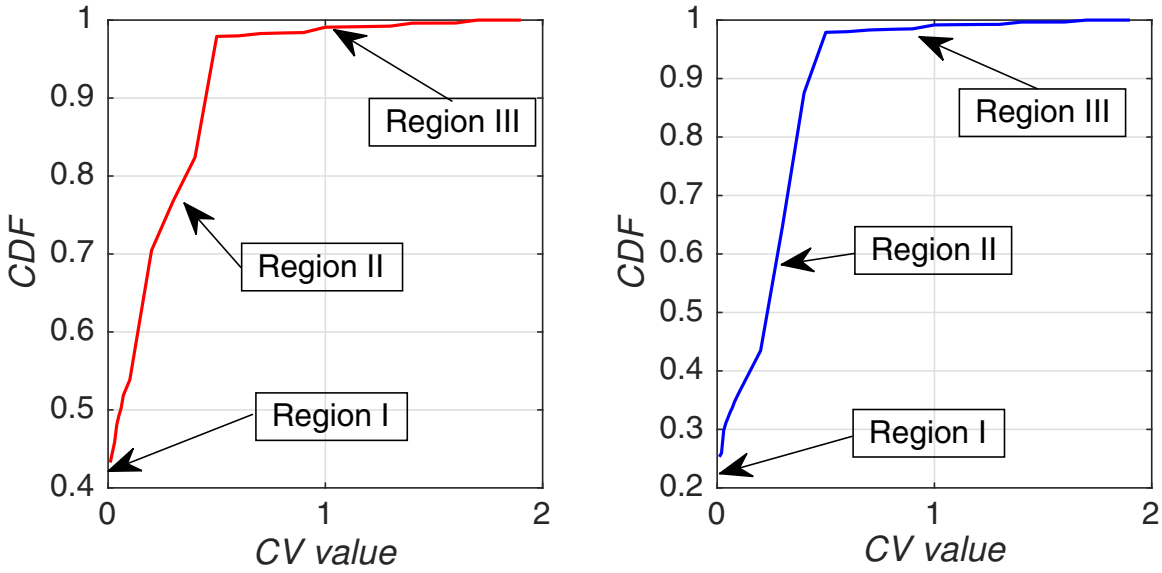
Therefore, we are using metric CV as the ratio standard deviation (σ) to the mean (μ) (see Equation (3.1)). When CV of a block equals one, then the depth values of

that block have the same mean and standard deviation values. It can also be argued that blocks with large CV values are probably related to several objects or edges while blocks with very small CV values are related to the background of the video frame. Consequently, they are not the interesting part of the frame that the RoI extraction algorithm is looking for.



(a) Video "Interview" (b) Video "Orbi"

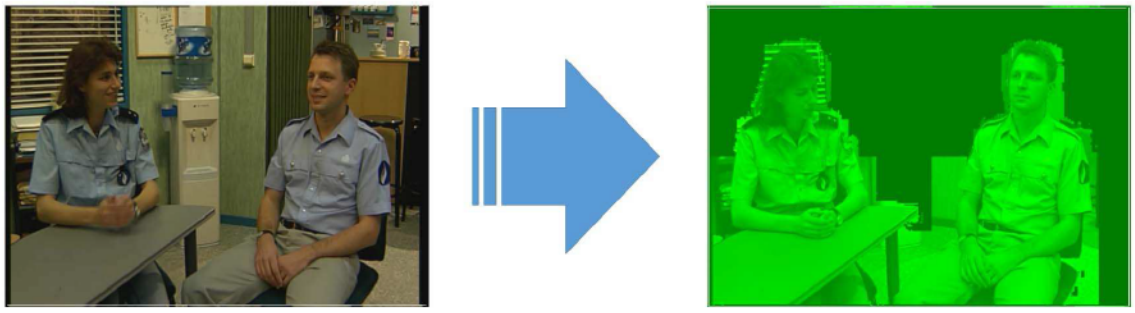
Figure 3.6: PDF of CV for depth map image (frame 1).



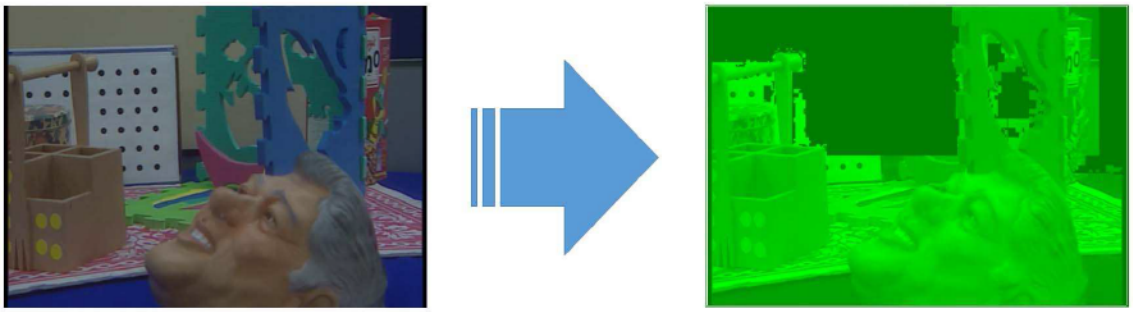
(a) Video "Interview" **(b)** Video "Orbi"

Figure 3.7: CDF of CV for depth map image (frame 1).

The typical probability density function (PDF) and cumulative density function (CDF) of CV values for videos "Interview" and "Orbi" are shown in Figures 3.6 and 3.7, respectively. Like Figures 3.4 and 3.5, the same argument is applicable for PDF and CDF as shown in Figures 3.6 and 3.7. A sample of map detection for region II (interesting objects) using the CV metric is shown in Figure 3.8.



(a) Video "Interview"



(b) Video "Orbi"

Figure 3.8: Example of region II detected by CV.

As can be seen in this figure, the identified RoI with CV values is considerably more accurate than the similar region shown in Figure 3.2. Compared to Figure 3.2, there is no missed block inside of interesting objects, nor does the background include interesting objects. More results for comparison between the performance of PV metric and CV metric will be presented in Section 4.2.

3.2 MDC Algorithm

This thesis proposes two multiple description coding methods focusing on region of interest (RoI) of the frame. These methods which are called spatial multiple description coding and spatiotemporal multiple description coding, are explained in Section 3.2.1 and Section 3.2.2, respectively.

3.2.1 Spatial Multiple Description Coding Focusing on RoI

Figure 3.9 shows the block diagrams of the first MDC method, i.e. spatial MDC method. As labelled on the figure, this method includes three parts: RoI Extraction

Algorithm, MDC Polyphase Subsampling Algorithm, and Descriptions Enhancement Algorithm. Each part will be explained in detail in Section 3.2.1.1, Section 3.2.1.2, and Section 3.2.1.3.

In summary, first we split the depth map image and extract the map for regions I, II, and III with the help of the algorithm explained in Section 3.2.1.1. Then four symmetrical descriptions are created using a polyphase sub-sampling algorithm as explained in Section 3.2.1.2. Afterward, as explained in Section 3.2.1.3 the descriptions are modified so that the ROI and the edges regions contain more information. The performance of this method will be examined in the next chapter.

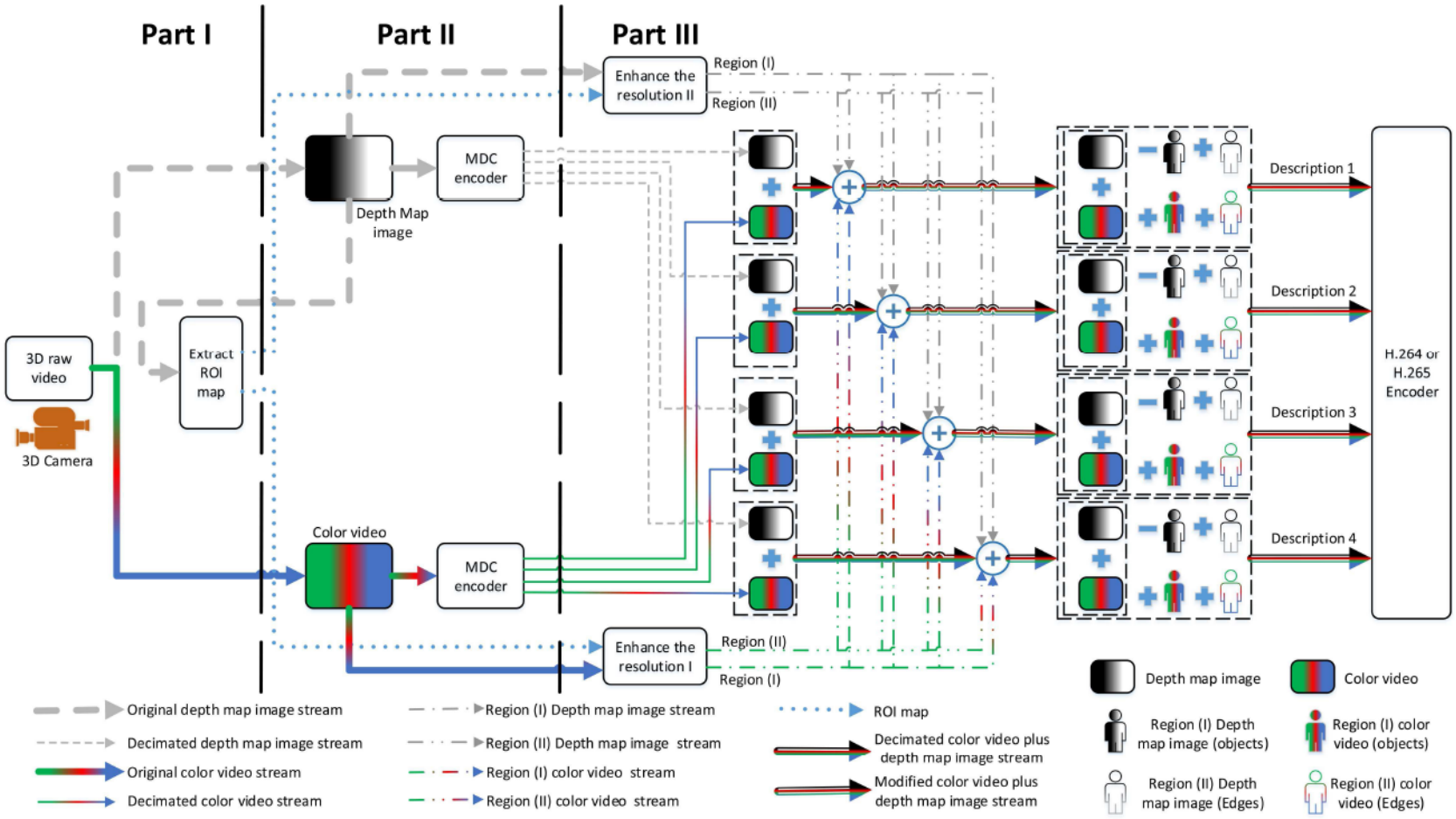


Figure 3.9: Block diagram of the spatial MDC method.

3.2.1.1 RoI Extraction Algorithm

As shown in Figure 3.9, the first step of the proposed encoder is to determine which parts of the frame are more important. The key requirement of this process is that it must have a low complexity algorithm to detect the interesting objects in the frame. A depth map can be used to extract the RoI map because it includes low frequency content compared to the color video frames that contain such a variety of frequency components.The depth map image consists of two major components:

- Low-frequency contents: the depth information for natural objects is usually similar, the depth map image mainly contains low-frequency contents.
- Edges: depth maps usually show sharp edges due to the different depth information of foreground and background objects.

The RoI extraction algorithm uses the characteristics of the depth map image and extracts the map of RoI using one of the metrics explained in the previous section.

Figure 3.10 shows the hierarchical block division (HBD) algorithm used by this thesis to identify the objects. The RoI range is defined as the distance between σ_{min} and σ_{max}, in this figure. σ_{min} is the threshold which is used to separate the very far objects in the background from the interesting objects and σ_{max} is the limit used to detect edges of the interesting objects. Clearly, the RoI range is different for two metrics that were introduced in Section 3.1 and shown by $[\sigma_{min}^{PV}, \sigma_{max}^{PV}]$ and $[\sigma_{min}^{CV}, \sigma_{max}^{CV}]$. The minimum thresholds are set so that the distant background can be separated from the interesting objects. Based on our experiment results in Figures 3.4 to 3.7, σ_{min}^{PV} and σ_{min}^{CV} can be a value approximately between [0.1 0.3] and [0.01 0.1], respectively. The maximum thresholds are selected to separate the interesting objects from their edges. As can be seen in Figures 3.4 to 3.7, σ_{max}^{PV} and σ_{max}^{CV} can also be a value approximately between [1 3] and [0.5 1], respectively. It is worth mentioning that N_{itr}^{Tot} is the total possible number of iterations that can be run by the hierarchical block division algorithm.

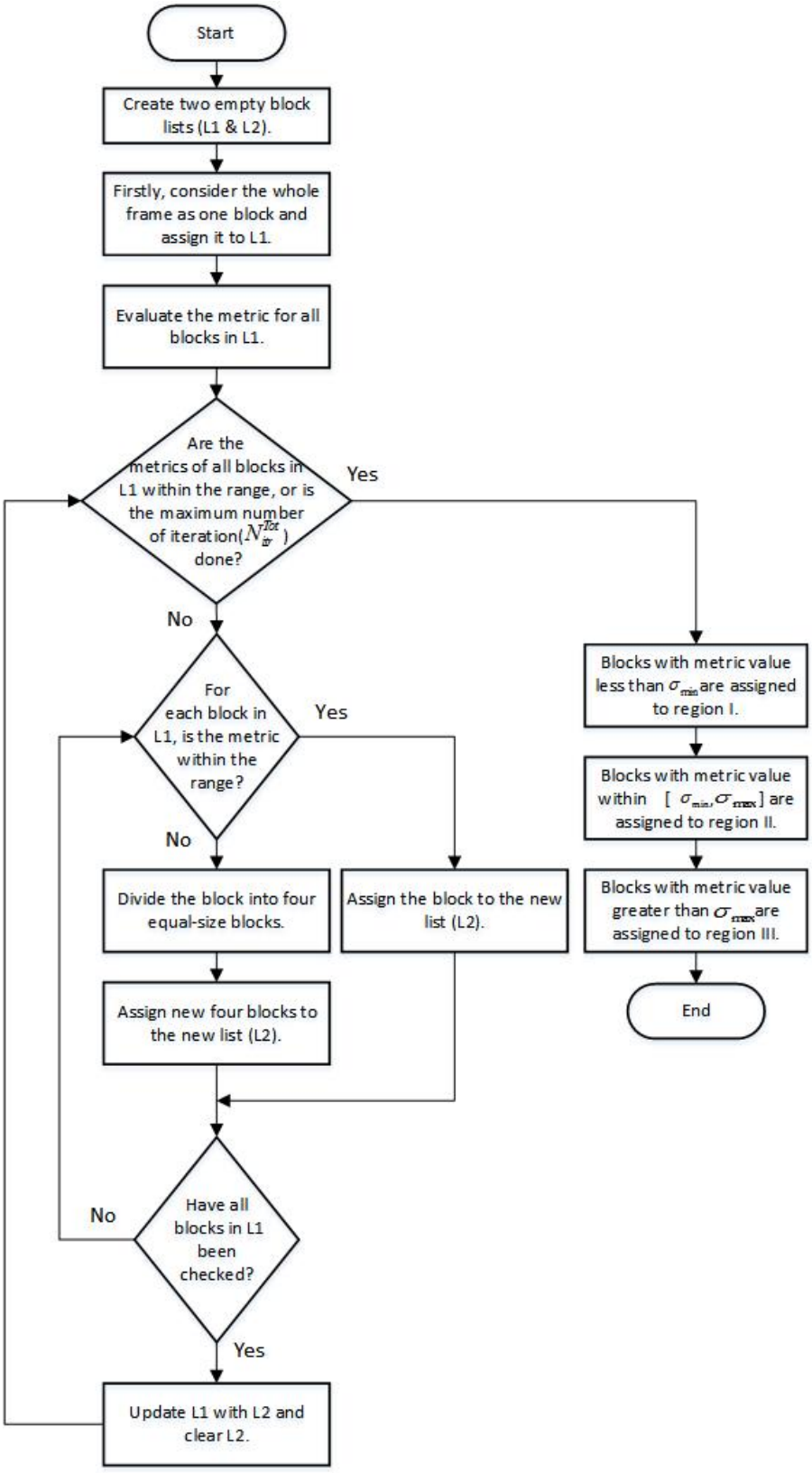


Figure 3.10: The algorithm of identifying important pixels.

Figure 3.10 illustrates the four major steps of this algorithm:

- **Step 1:** Create two empty lists (L_1&L_2), and assign the entire depth map image as one block to L_1. Then start the first iteration as explained in step 2.
- **Step 2:** Check if the algorithm reaches the limit of N_{itr}^{Tot} or if all blocks in L_1 are with PV or CV values smaller than σ_{max}^{PV} or σ_{max}^{CV}, respectively. If yes, go to step 4. If not, go to step 3. Clearly, in the first iteration there is only one block in L_1 and its metrics are with the strong probability greater than σ_{max}.
- **Step 3:** For every block in L_1 with the metric value greater than the threshold, divide the block into four equal sized blocks and assign them to L_2. Any block with metric value less than the threshold is assigned without change to L_2. After having checked all the blocks in L_1, L_1 is updated with L_2 and L_2 is cleared. Then return back to the second step.
- **Step 4:** All blocks in L_1 with metric values less than σ_{min} are considered as region I. Blocks with metric values within the RoI range are considered as region II and remaining blocks are classified as region III .

In the hierarchical block division algorithm, a block is partitioned into smaller blocks by dividing the width and height of the block by a factor 2 in each iteration. It is worth mentioning that N_{itr}^{Tot} should be defined in order that the minimum block size be greater than a 2×1 or 1×2 pixels block size. This is due to the fact that both metrics used in this algorithm evaluate pixel variation where there is at least two pixels to measure the variation.

A sample hierarchical block division process is shown in Figure 3.11. You may assume that the numbers inside the blocks represent typical values for the metric values used in the algorithm. For this figure, it also has been assumed that the resolution of the depth map image is 16×16 pixels and the smallest block is of the size 2×2. Clearly, N_{itr}^{Tot} is 4 and σ_{max} can be assumed as 10. The highlighted blocks in the fourth iteration are showing the important region of the frame that the proposed algorithm is looking for. As can be seen in this example, there are some blocks with large metric values resulting in further partitioning but because the number of algorithm iteration reaches N_{itr}^{Tot}, the algorithm stops partitioning blocks. In this example, σ_{min} was not defined, but in practice this parameter should be used

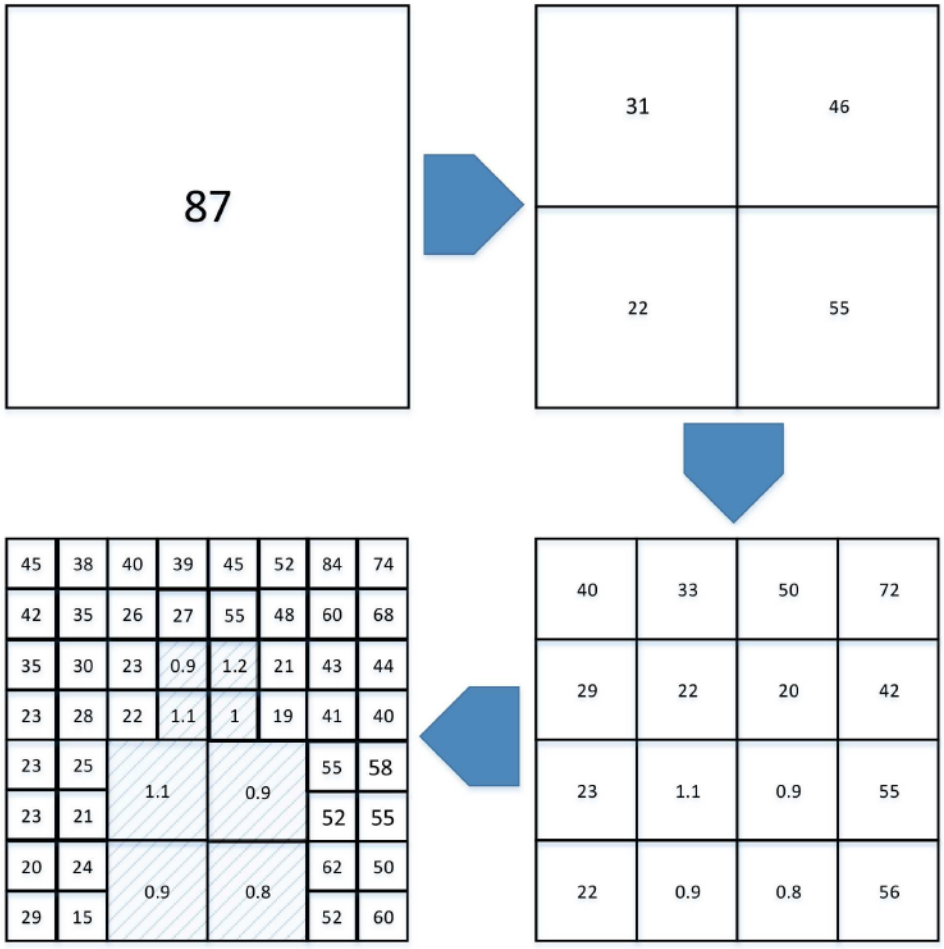


Figure 3.11: Hierarchical division process to identify RoI with $\sigma_{max} = 10$.

in order to separate very far background objects with very small depth values (close to zero) from interesting objects in the depth map image. Because the background region is often out of focus during the capturing of videos, this background region also needs to be excluded from RoI in the proposed algorithm. Performances of the RoI extraction algorithm for the metrics PV and CV are examined and compared in Chapter 4.

3.2.1.2 MDC Polyphase Subsampling Algorithm

To have reliable video streaming, the proposed new spatial MDC algorithm exploits the Multiple Description Coding (MDC) strategy for 3D videos. To this end, four

descriptions are created using Poly phase SubSampling (PSS). PSS-MDC is the basic low complex method that can be used in the spatial domain to have a reliable transmission in the error prone environment. As can be seen in Figure 3.12, with the PSS-MDC encoder used by the proposed method every description includes one of 2×2 pixels. Since the new spatial MDC algorithm is applied on 3D stereoscopic videos, the PSS-MDC encoder needs to be applied on both color and depth map frame separately.

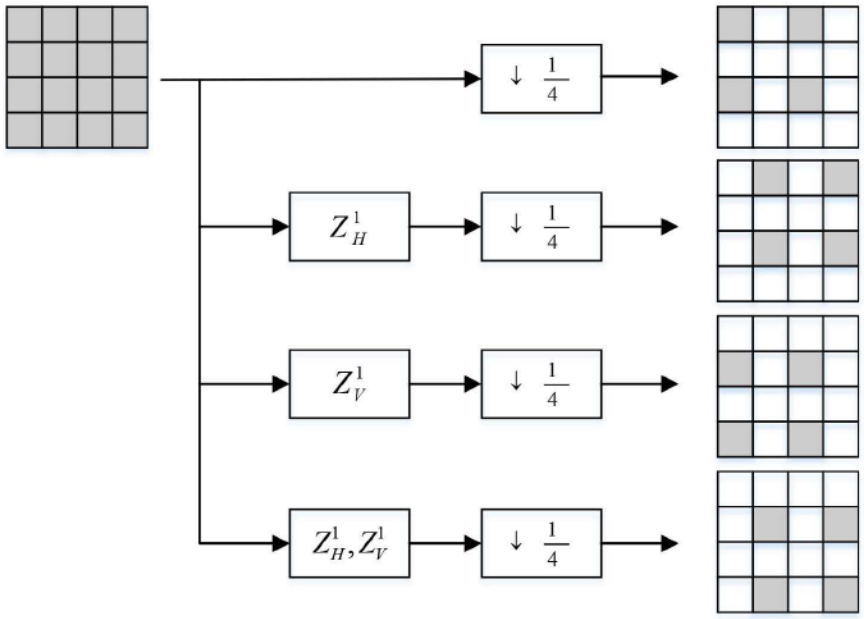


Figure 3.12: Polyphase SubSampling MDC encoder used in Figure 3.9 (Z_H^1 and Z_V^1 are horizontal and vertical shift, respectively).

3.2.1.3 Spatial Descriptions Enhancement Algorithm

Although the most important advantage of the PSS-MDC encoder is its simplicity, there is no adjustment tools to increase the redundancy to avoid errors in the strong noisy environment. To fix this, the new spatial MDC algorithm enhances the spatial resolution for areas that are less predictable and also on objects of interest that are more important to focus on.

The section labelled as Part III in Figure 3.9 shows the block diagram of spatial descriptions enhancement algorithm. As can be seen, two different algorithms are applied to the color video and the depth map stream. For the depth map stream, the

resolution of each description is enhanced according to its prediction difficulty. Since the metrics defined in this thesis evaluate the variation between adjacent pixels, it can be said that pixels of the depth map frame are clustered into regions I to III according to their difficulty prediction levels. This means that the region I, which includes pixels with very low variations, remains without any change. Pixel resolution in the region II is enhanced to one second for each description in the encoder by picking any pixels of 2×2 pixels other than the pixel was initially assigned to the description. Since it has been assumed that three descriptions are lost due to the unreliable communication and only one description is available in the decoder, it is of minor importance that which pixel is added to the pixels in the one forth resolution. Since the region III contains pixels with large variations, it is likely that the prediction of a pixel (in case of missing) from adjacent pixels leads to error. As a result, this region's pixel resolution has increased to a fuller pixel resolution for each description.

Since the region's clustering algorithm is done using the depth map image rather than the color video frame, it cannot reflect the pixels' value variations for the color video frame. Therefore, the above-mentioned argument is no longer applicable. One suggestion with regards to the color video is to apply the proposed RoI detection algorithm on the color video stream in order for it to extract RoI map based on the pixel variation found in the color video frame; but the drawback is its greater complexity due to a wide variety of colors inherently part of any scene naturally. As a result, the hierarchical block division algorithm needs more time to identify different regions in the frame. Another suggestion is to use the RoI map extracted from the depth map image to then focus on region II for the enhancement of pixel resolution in the color video frame rather than on region III which is performed within the depth map stream. Since the human eye is more sensitive to objects rather than of pixels, this suggestion introduces better performance with regards to the subjective assessment. Also, it can provide improvement with regards to the objective assessment since the recording of moving objects which are inherently part of the scene, are now more focused. Because all video coding standards use differential pulse code modulation (DPCM) and proximate pixels' values of the objects in the color video frame, the increase of the resolution of those parts of a frame that include the RoI can be compensated by DPCM algorithm in point of compression ratio. Therefore, with regards to the color video stream, region II and III are enhanced to full and one-second resolution, respectively. Region I remains with the same resolution

as before (one fourth). This enhancement algorithm helps to perfectly recover the RoI in the instance of missing a description, although at the expense of increased redundancy.

3.2.2 Spatiotemporal Multiple Description Coding Focusing on RoI

Figure 3.13 shows the block diagrams of the hybrid MDC method, i.e. the spatiotemporal MDC method. As with the spatial MDC Algorithm presented previously, it includes three primary parts: the RoI Extraction Algorithm, the MDC Polyphase Subsampling Algorithm, and the Descriptions Enhancement Algorithm. The first two parts are the same as before (for more details see Section 3.2.1.1 and Section 3.2.1.2). Only part III, the description enhancement algorithm, differs. As opposed to the previous algorithm, which gave more bandwidth to regions II and III, this algorithm decreases the bandwidth assigned to region I and increases it to regions II and III. Section 3.2.2.1 describes this method in more detail. The performance of this method will be examined in the next chapter.

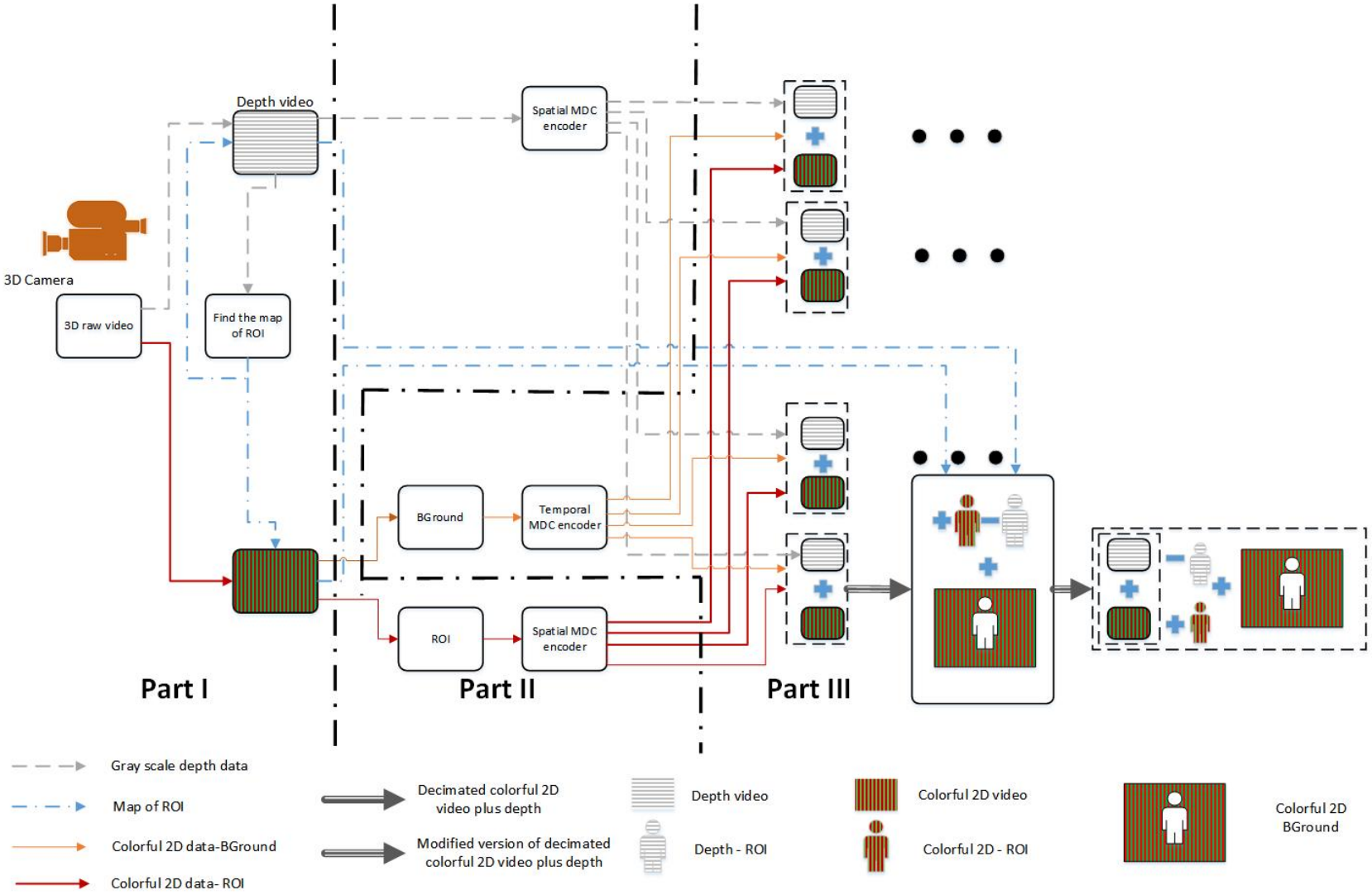


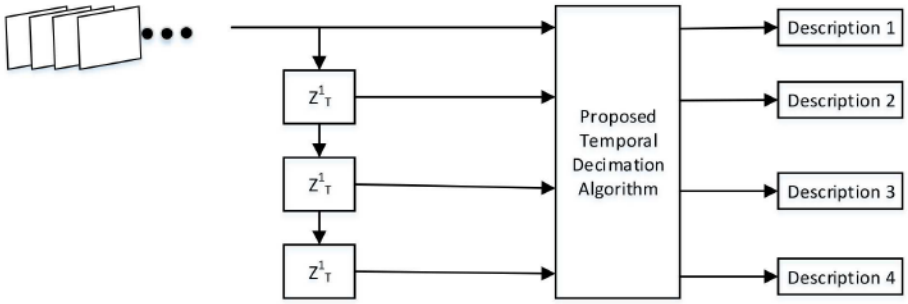
Figure 3.13: Block diagram of the spatiotemporal MDC method.

3.2.2.1 Spatiotemporal Descriptions Enhancement Algorithm

The spatial descriptions enhancement algorithm assigns more bandwidth to the ROI and edges than to the background, as discussed in the previous section. Therefore after reconstructing the video at receiver, only main objects are displayed with high quality and background objects are shown with a lower quality. It is worth mentioning that background is the part of a frame that usually has slow movement. Therefore a missed pixel can be predicted more accurately from the pixel located at the same position from the adjacent frames. Thus instead of looking at surrounding pixels at the same frame, it is better to look for a similar pixel at the adjacent frames. To compensate for the redundancy added to each description for the RoI part, the encoder drops every other background (or I in this thesis). As mentioned earlier, Figure 3.13 shows the block diagram of spatiotemporal MDC method.

Figure 3.14 shows how the temporal modification process augments each description information. As shown in Figure 3.14a for every four frames, each description increases the resolution of background (or not RoI part) to the full resolution. For example, in description one, only frames $1, 5, ...$ have complete resolution information of the background. To avoid a huge increase in the volume of data in each description, the first and third frames after the full resolution frame are completely dropped from each description and they are predicted from the adjacent frames. Because of the low movement of background, the missed frame can be reconstructed with higher quality.

(a) Temporal decimation algorithm



(b) Temporal multiple description coding

Figure 3.14: Temporal modification process: Figure 3.14a shows the temporal decimation process (green: increase the resolution to full resolution after MDC decimation, yellow: without change, red: dropped) and Figure 3.14b describes how descriptions are created using a temporal MDC algorithm (Z^1_T is time shift or next frame).

It should be noted as the proposed method target the live stream application and time is crucial, decoder looks for the missed pixels in the previous frame.

3.3 Conclusion

Multiple description coding (MDC) is a solution against packet failure when multimedia are streamed through the network in error prone environment. With MDC, one video description is partitioned into several separately decodable descriptions. In the instance of missing a description during transmission, the decoder is capable to estimate the lost description from other error free description(s). To improve the basic spatial partitioning and to be applicable to 3D videos, a non identical decimation algorithm for the stereoscopic videos has been provided. Our algorithm works based on existing objects in the scene and assigns more bandwidth to the region of

interest. It is worth mentioning that they are many neural network or learning algorithms to extract objects; however, learning algorithms relies on training. Such algorithms need to be trained first to be able to recognize specific objects and in case of having object other than trained objects it fails. Unlike the learning algorithms the provided algorithm does not require training and is applicable on any stereoscopic video. In addition to that, one restriction for learning algorithms is complexity to consider many scenarios and choose the best solution for the objective function. Such property, make learning algorithms unpractical for live stream application in which time is very crucial.

Chapter 4

Experiments and Results

4.1 Test Setup Information

The purpose of this chapter is to explain the simulation tests used to evaluate the proposed MDC methods. First, we explain how tests were conducted and what is the goal of the tests. Then we introduce the video test sequences used in our simulation. We will then discuss the results.

4.1.1 Test Scenarios

In this thesis we conducted two test scenarios in general as described in Section 4.1.1.1 and Section 4.1.1.2.

4.1.1.1 Test Scenario One

We are testing how missing descriptions affects the reconstruction of video at the receiver. In this scenario, we consider the worst-case scenario with the least amount of description available at the receiver. The steps to perform this test scenario are as follows:

1. We first need to split the depth map image from the color image as the input video is in YUVA format and its chroma and depth subsampling format is 4:2:2:4.

2. The RoI map image is then extracted from the depth map image using the algorithm explained in Section 3.2.1.1. As explained in Section 3.2.1.1, we are

extracting the RoI map using two metrics PV and CV. The result is also compared in Section 4.2.

3. Four descriptions are created with a poly phase subsampling algorithm as explained in Section 3.2.1.2. In order to assign more bandwidth to areas of interest, descriptions need to be modified. As explained in Section 3.2.1.3 and Section 3.2.2.1, descriptions are modified in two domains: "spatial" and "spatiotemporal".

4. The descriptions now must be encoded. In this thesis, we are using H.264/AVC reference software (JM 19.0) [146] and H.265/HEVC reference software (HM 6.0) [147] to encode the test videos. To encode with reference software, I frames (or independent reference frames) are repeated every 4 frames and only P (or predicted frame) frames are used between I frames. It is worth mentioning that predicted frame holds only the changes in the image from the previous frame. To simulate this scenario for different rates, we changed the Q parameter at the reference software from 20 to 40 in increment of 5. (Q parameter is used for the quantization. Generally how smaller the Q is, the lower quantization error and better quality are.)

5. In order to consider an error-prone environment, we assume that there is only one description available at the receiver. Since the descriptions are symmetrical, it does not matter which description is available at the receiver. Picking the first description, we interpolate missed pixels using neighboring pixels. The results will be examined in Section 4.3.1.

The objective in this scenario is to examine the robustness of the proposed method if three descriptions are completely lost. We expect to have a higher quality of reconstruction compared video decoded by the normal MDC (without focusing RoI pat). When the proposed hybrid MDC is used, the improvement is even greater.

4.1.1.2 Test Scenario Two

We are simulating random packet failure in this scenario, which is common in network communication. For this purpose, we segment and encode all descriptions with I frames for every 4 frames and the Q parameters set to 30. Each encoded segment is then divided into 1500-byte packets. The packets are dropped at a random rate of

2.5% and 5% to simulate a noisy channel or poor network transmission quality. A drop rate of more than 2.5% indicates poor network quality. The details of this system will be discussed in Section 4.3.2. Then, we reconstruct the video from the available packets. This scenario is intended to determine whether the proposed method can actually be implemented. For this scenario, we also expect to have a better result however it is not as much as the previous scenario as packets of all descriptions may be lost and there is no available description for that segment to do the estimation. In this case the decoder estimate the lost data from the same position of the previous block.

4.1.2 Test Video Sequences

The following three types of 3D video were used in this thesis:

1. A stereoscopic video sequence called "Interview" and a stereoscopic video sequence called "Orbi" were originally distributed in Standard Definition (SD) TV format (576×720) at a framerate of 25 frames/second. Using a depth map camera, these test sequences produce clean depth map frames. "Orbi" is a very complex test video with multiple objects at different depths and camera movements. Video "Interview" is a motionless background and fixed camera sequence, which simulates video conference scenarios [2, 148]. The chroma and depth subsampling format is 4:2:2:4 (The first 4 refers to the size of the sample. The two following 2 both refer to chroma. They are both relative to the first number and define the horizontal and vertical sampling respectively. The last 4 indicates the resolution of the depth map). There are also 90 frames in each sequence.

2. Two stereoscopic video sequences called "Break dance and "Ballet and are initially in extended graphics array (XGA) format (768×1024) and 15 frames/s frame rate. Both sequences are captured with static multi-view cameras generated by the Interactive Visual Media group at Microsoft Research [149]. This thesis utilizes the fourth camera view and related depth map computed from stereo-views. In the video "Break dance", one object moves rapidly in the foreground and several objects remain motionless in the background. Contrary to the video "Break dance, the video "Ballet contains a stationary object in the foreground and the main non-stationary object is located behind it [2].

Chromatic and depth subsampling formats are 4:2:2:4, and the total number of frames is 90.

3. A video sequence called "Pingpong" created as part of this thesis at Carleton University, originally had the resolution of (1024×2048) and a frame rate of 30 frames/s. As JM software does not support resolutions more than 1K, the video was decimated to 512×1024. The depth map frame was generated using depth map automatic generator (DMAG) [150] and the video was captured using two fixed-position cameras. like the other two types, the chroma and depth subsampling format is 4:2:2:4, and total number of frames is 150.

The hierarchical block division algorithm described in Chapter 3 is used to extract ROI from a frame by first halving it in both width and height dimensions, making smaller blocks with depth changes less than a threshold.

As described in Chapter 3, we are using the hierarchical block division algorithm to extract RoI of a frame, first. To this end, a frame is halved hierarchically in both width and height dimensions to make smaller blocks with depth change less than a threshold. It should be noted that for "Interview" and "Orbi" video sequences, the width of the depth map frame, which is 720, is not divisible after the fourth iteration (as 720 equals $2^4 \times 3^2 \times 5$). To be able to continue the hierarchical block division algorithm after the fourth iteration, we extend the depth width to 768 ($= 2^8 \times 3$) by adding zeros to the left side of the depth map image. With the same argument, the height of the depth map frame is assumed to be 512 ($= 2^8 \times 2$). Therefore, the acceptable minimum size of a block at the end of the hierarchical division algorithm is 2×3 which is achieved after the eighth iteration (this means that the N_{itr}^{Tot} equals 8 and the minimum block size after the eighth iteration of the hierarchical block division algorithm is 6 pixels).

The minimum and maximum thresholds used for the tests presented in this thesis are $[\sigma_{min}^{PV} = 0.3, \sigma_{max}^{PV} = 3]$ and $[\sigma_{min}^{CV} = 0.01, \sigma_{max}^{CV} = 0.75]$. In the remainder of this thesis, we will first discuss the complexity of the proposed algorithm in general. Then, both the performance and complexity of the proposed MDC methods using CV and also PV will be compared. Thereafter, we will evaluate the performance of the new proposed spatial MDC algorithm for streaming in the error prone environment.

It is worth mentioning that in this thesis we used $PSNR$ (Peak signal-to-noise ratio) and $SSIM$ (structural similarity index) for assessment purpose. PSNR is an technical term for the ratio between the maximum possible value of a image and the

power of corrupting noise and is usually expressed as a logarithmic quantity using the decibel scale. PSNR is commonly used to quantify reconstruction quality for images and video subject to lossy compression. SSIM is a term for predicting the perceived quality of video and is used for measuring the similarity between two images. These two metric are chosen in this thesis as they are the most common methods and also to have a compassion between our method with similar works.

4.2 Performance Examination of Metrics

The ROI of the frame is calculated using two metrics described in Section 3.1. We run our block division algorithm on the depth map image to cluster a frame to three regions: from Region I with a lower depth change to Region III with a higher depth change. Figure 4.1 shows the identified Regions I to III using PV and CV metrics. With the CV metric, the identified Region II is clearly more accurate than with the PV metric. The same scenario is also applicable to the Region I. As can be seen in part (d) of Figure 4.1 there are some important pixels that have not been identified as the region II (RoI). Also we have detected some missed pixels in region I (background) with PV as shown in part (b) of Figure 4.1. An inaccuracy in identifying different regions with PV can be attributed to the fact that the pixel values of different blocks are in different ranges. Therefore the pixel variation (PV) cannot be an appropriate metric to be used when extracting for regions I and II. To fix this problem as argued before, it is necessary to normalize the pixel variation metric (PV). Indeed, the CV metric is the normalized pixel variation and works like a smoothing filter. Although using normalized pixel variation metric (CV) provides a considerable improvement in the extraction of regions I and II, such performance is not shown when using the CV metric in identifying region III (which stands for the edges). As can be seen in Figure 4.1, the detected edges shown in part (g) is not as clear as the detected edges shown in part (f). The reason for that can be the smoothing effect brought about by the normalization using the CV metric. As the blocks that contain edges are considered as blocks with high-frequency contents, a high-frequency filter like the pixel variation measurement (PV) is more beneficial for identifying the edges.

(a) Original video frame.

(b) Extracted region I by PV.

(c) Extracted region I by CV.

(d) Extracted region II by PV.

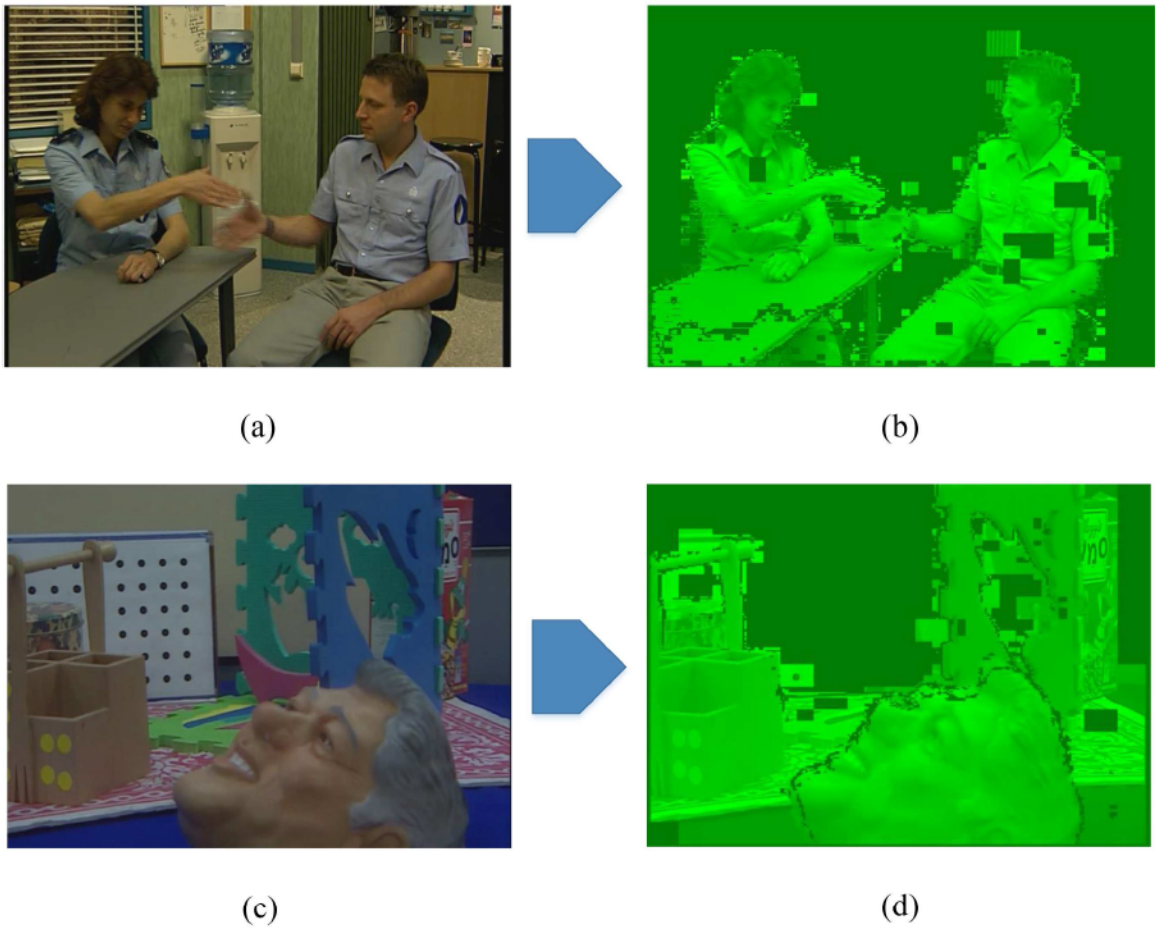
(e) Extracted region II by CV.

(f) Extracted region III by PV.

(g) Extracted region III by CV.

Figure 4.1: Performance comparison between identified regions I-III for the first frame of video "Interview".

Figure 4.2 shows the original 2D video frames (the 87th frame of video "Interview" and the 1st frame of video "Orbi"), as well as their identified ROIs. Detecting pixels related to the hands of objects in the video "Interview" during handshaking (moving objects) can show the acceptable performance of the algorithm for realizing RoI. This figure also shows a good performance for the second test sequence, i.e. video "Orbi". As can be seen objects in this video have been identified very well by the proposed algorithm.



(a) (b)

(c) (d)

Figure 4.2: A sample performance of RoI identification using metric PV: (a) Original 2D video (video "Interview"). (b) Detected RoI (video "Interview"). (c) Original 2D video (video "Orbi"). (d) Detected RoI (video "Orbi").

Table 4.1 shows the average number of blocks in different sizes after the hierarchical division algorithm for the videos "Interview" and "Orbi". As can be seen, there is one block with the size of 24576 ($= 128 \times 192$) using PV metric and 4 blocks of

this size using CV metric for test video Interview. For video Orbi, there are only two blocks of this size using PV metric or CV metric. The reason that these numbers are equal is we have several objects in video Orbi. This causes difficulty to detect objects and both metrics work with the same performance. It is worth mentioning this table only includes the data related to the first frame. This means that about $5 \sim 9\%$ of the entire depth map image is excluded from being more partitioned and stopped after the second iteration of the hierarchical division algorithm. Considering the second large block size for PV metric in Table 4.1, i.e. 6144 ($= 64 \times 96$), it can be said that the hierarchical division process will be stopped for more than one-third of the entire depth map image in the video Interview and one fourth of the depth map image in video Orbi after the third iteration. This result shows that the hierarchical division algorithm does not give rise to a high load of calculation in this proposed algorithm. Using the CV metric results in a decrease in the complexity level as compared to the PV metric. As can be seen the largest block size for the CV metric is increased to 98304 ($= 256 \times 384$) and the average number of blocks with the size of 24576 is two in the video Interview and three in video Orbi for metric CV.

Table 4.1: Number of blocks with different sizes after hierarchical division algorithm using metrics PV and CVb (first frame only).

(a) Video "Interview".

	Metric PV		Metric CV	
Blocks' size	No. Blocks	percent	No. Blocks	percent
6 (2×3)	3401	5	2168	3
24 (4×6)	995	6	358	2
96 (8×16)	509	12	169	4
384 (16×24)	210	20	81	8
1536 (32×48)	48	19	47	18
6144 (64×96)	21	32	25	40
24576 (128×192)	1	5	4	23
98304 (256×384)	0	0	0	0

(b) Video "Orbi".

	Metric PV		Metric CV	
Blocks' size	No. Blocks	percent	No. Blocks	percent
6 (2×3)	5254	8	2264	3
24 (4×6)	1056	6	378	2
96 (8×16)	505	12	160	4
384 (16×24)	177	17	72	7
1536 (32×48)	74	29	52	20
6144 (64×96)	11	17	13	20
24576 (128×192)	2	9	2	10
98304 (256×384)	0	0	1	33

Table 4.2 shows the average number of blocks for different metric values of PV and CV. As can be seen, about 55% of the depth map image for video Interview and 40% of the depth map image for the video Orbi have PV values less than 1. In other words, for the video Interview more than one half and for video Orbi more

Table 4.2: Number of blocks with different metric values after hierarchical division algorithm.

(a) Video "Interview".

		Blocks' size								Percent of blocks with metric value in a specific range(%)
		6 (2×3)	24 (4×6)	96 (8×16)	384 (16×24)	1536 (32×48)	6144 (64×96)	24576 (128×192)	98304 (256×384)	
PV	≤ 1	663	372	173	76	22	18	0	0	55.68
	$1 \sim 3$	1008	618	336	134	25	3	0	0	41.64
	$3 \sim 10$	832	5	0	0	0	0	0	0	1.30
	≥ 10	899	0	0	0	0	0	0	0	1.37
CV	≤ 0.1	646	277	151	68	38	19	0	0	56.74
	$0.1 \sim 0.2$	33	11	4	3	3	1	1	0	15.57
	$0.2 \sim 0.3$	45	16	6	4	4	2	0	0	5.96
	$0.3 \sim 0.4$	105	25	4	3	2	2	0	0	5.22
	$0.4 \sim 0.5$	53	29	4	4	0	0	0	0	14.55
	≥ 0.5	1286	0	0	0	0	0	0	0	1.96

(b) Video "Orbi".

		Blocks' size								Percent of blocks with metric value in a specific range(%)
		6 (2×3)	24 (4×6)	96 (8×16)	384 (16×24)	1536 (32×48)	6144 (64×96)	24576 (128×192)	98304 (256×384)	
PV	≤ 1	543	295	173	69	34	5	0	0	39.37
	$1 \sim 3$	1681	753	332	108	39	6	0	0	55.95
	$3 \sim 10$	2276	8	0	0	0	0	0	0	3.53
	≥ 10	754	0	0	0	0	0	0	0	1.15
CV	≤ 0.1	614	245	119	49	35	7	0	0	39.28
	$0.1 \sim 0.2$	59	28	12	9	6	2	0	0	6.76
	$0.2 \sim 0.3$	80	28	9	6	4	3	0	0	19.80
	$0.3 \sim 0.4$	135	35	10	5	4	1	0	0	21.54
	$0.4 \sim 0.5$	90	42	10	4	2	0	0	0	10.66
	≥ 0.5	1286	0	0	0	0	0	0	0	1.96

than one-third of the depth map image have very close depth values. This is the reason why the decimation of the depth map image does not affect its quality when it is reconstructed in the decoder [127]. Table 4.2 also shows that about 95% of the depth map image for both test video sequences have PV values less than 3. The fact that about 95% of the depth map image have similar depth values result in no longer needing to send the depth map image with its original resolution, justifies why the nonidentical decimation is more advantageous than the identical decimation suggested by Karim et al. in [127].

Table 4.3 compares the statistics of the blocks generated by the hierarchical division algorithm using two metrics PV and CV. As shown in the table, the average

Table 4.3: Blocks specification

	Video	Interview	Orbi
Metric PV	Average Number of blocks per frame	5184.1	7078.47
	Average block size	4.08 × 6.12	3.51 × 5.27
	Average of PV values	6.86	4.3
Metric CV	Average Number of blocks per frame	2853.27	2942.2
	Average block size	4.22 × 6.33	3.91 × 5.86
	Average of CV values	0.548	0.539

block size for videos Interview and Orbi after hierarchical division algorithm are greater and the average number of blocks is considerably less when the CV metric is used. This means fewer operations are required to identify the final blocks using the CV metric. It also should be considered that the results obtained by the CV metric have a greater performance when compared to the results gained by the PV metric (see Figure 4.1). Therefore, better performance and less complexity can be achieved by using the new CV metric.

4.3 Error Resiliency Performance Examination

To show how robust the proposed method is against error, we are using PSNR (peak signal to noise ratio) and SSIM (structural index similarity) tools to evaluate the image quality at decoder. It is worth mentioning that PSNR and SSIM are two assessment tools widely used in multimedia applications to evaluate image quality. PSNR is older than SSIM and computes the mean squared reconstruction error. Typical PSNR result values are between 30 dB to 50 dB for compression, where higher means better quality. If the images significantly differ, PSNR is much lower, for example, 15 dB. In practice, PSNR may turn out inconsistent with human eye perception. The structural similarity algorithm aims to correct this. It has been developed to have a quality reconstruction metric that also takes into account the similarity of the edges (high-frequency content). In fact, it is designed based on luminance, contrast, and structure to better suit the workings of the human visionary system [151].

4.3.1 Test Results of Scenario One

Consider first a scenario in which the receiver only has one description. In this scenario, the decoder estimates the unavailable pixels at receiver using the nearest available pixel. Figure 4.3 and Figure 4.4 compare PSNR and SSIM measurements of the reconstructed color video for the video "Interview" using the basic poly phase subsampling MDC method (PSS-MDC, as most of new MDC method method using temporal MDC), the MDC method presented in [152], and the final proposed spatial MDC algorithm with the help of PV and CV metrics. As can be seen in Figure 4.3, compared to the work presented in [152] about 1 dB improvement with the PV metric and about 2 dB improvement with the CV metric can be achieved for the recreated video Interview. Our measurement shows that the improvement is more (about 4 dB) when it is compared to the Poly phase SubSampling MDC method. Such improvement comes from packing more related information (an object) in one description. Since an object's pixel information is more similar in comparison to several objects' pixels information, encoder compressed the description with more compression ratio. On the other hand, since we have more information related to the important objects, they are decoded with higher quality while not increasing the bandwidth considerably. In point of SSIM, the improvement is about 2% as shown in Figure 4.4. It is worth mentioning, that the most common MDC works is based on temporal MDC and therefore there is not enough works to have a fair comparison between our method with others'.

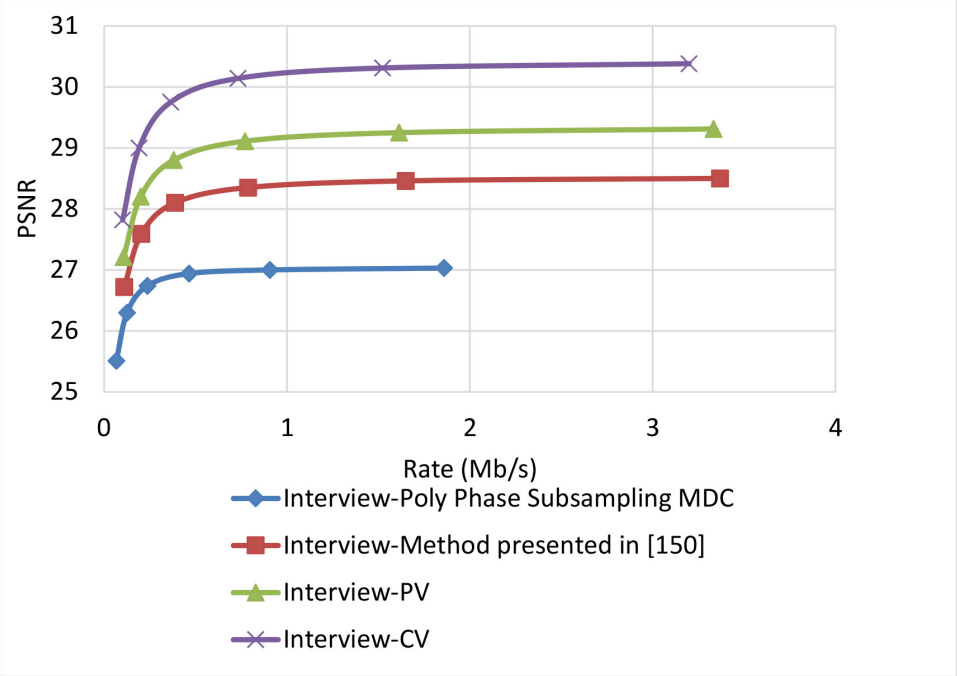


Figure 4.3: Average PSNR assessment of video "Interview" (color image).

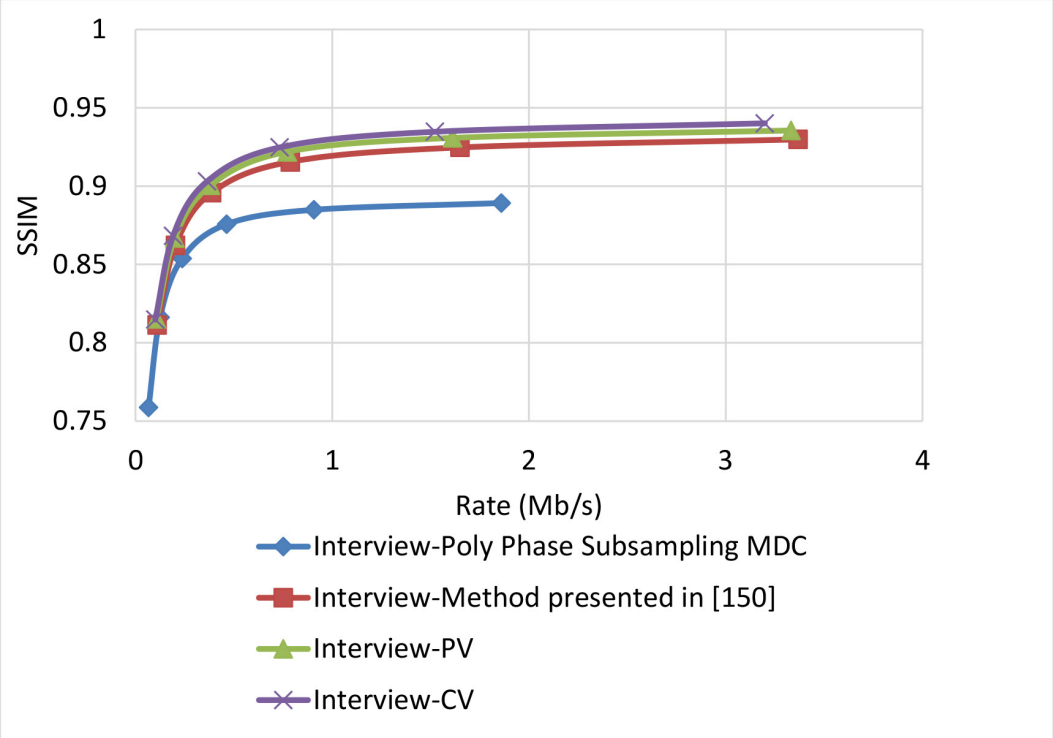


Figure 4.4: Average SSIM assessment of video "Interview" (color image).

Figure 4.5 and Figure 4.6 also show the same measurement for the video "Orbi". Although a considerable improvement cannot be seen compared to the previous work presented in [152], more than 2 dB improvement has been achieved by the new proposed spatial MDC algorithm in comparison with the PSS-MDC method. It should be noted that the simplicity of implementing the proposed method with CV is approximately doubled according to Table 4.3. In this table, the hierarchical block division algorithm for the PV metric is the same as the algorithm presented in [152]. Since the average number of blocks is about 0.5 and the average size of blocks is greater for the CV metric compared to the PV metric, we have concluded that the new proposed spatial MDC algorithm using the CV metric is more efficient. Regarding the SSIM assessment, the proposed algorithm provides about 3% improvement for both test videos in high rate streaming compared to the PSS-MDC method.

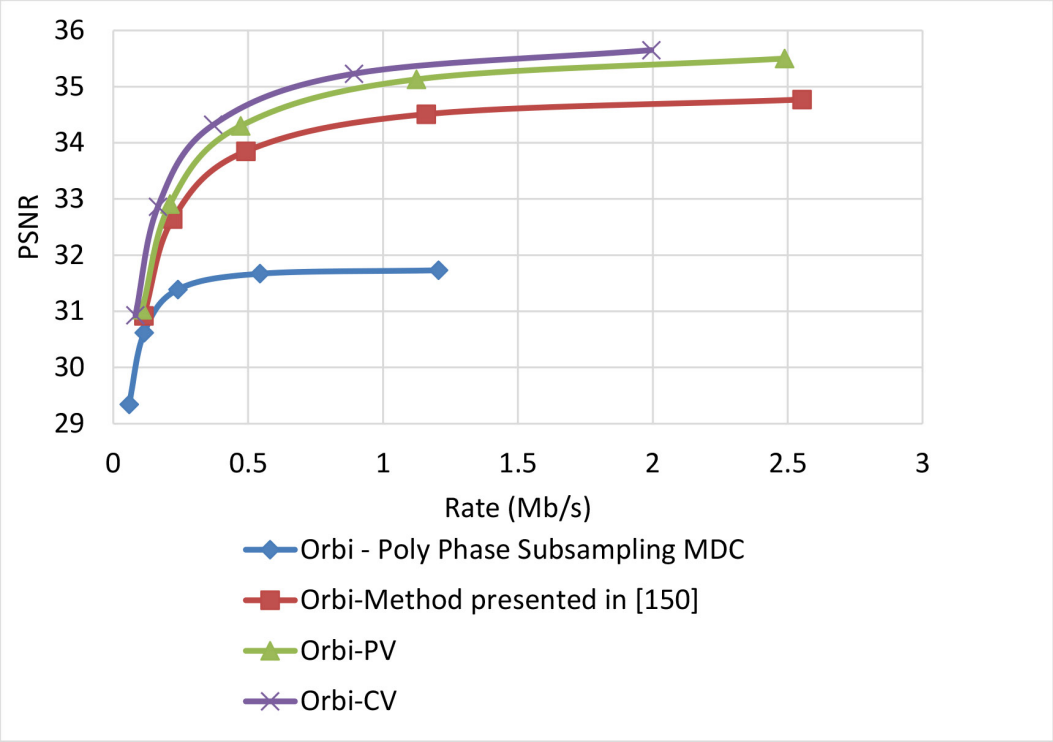


Figure 4.5: Average PSNR assessment of video "Orbi" (color image).

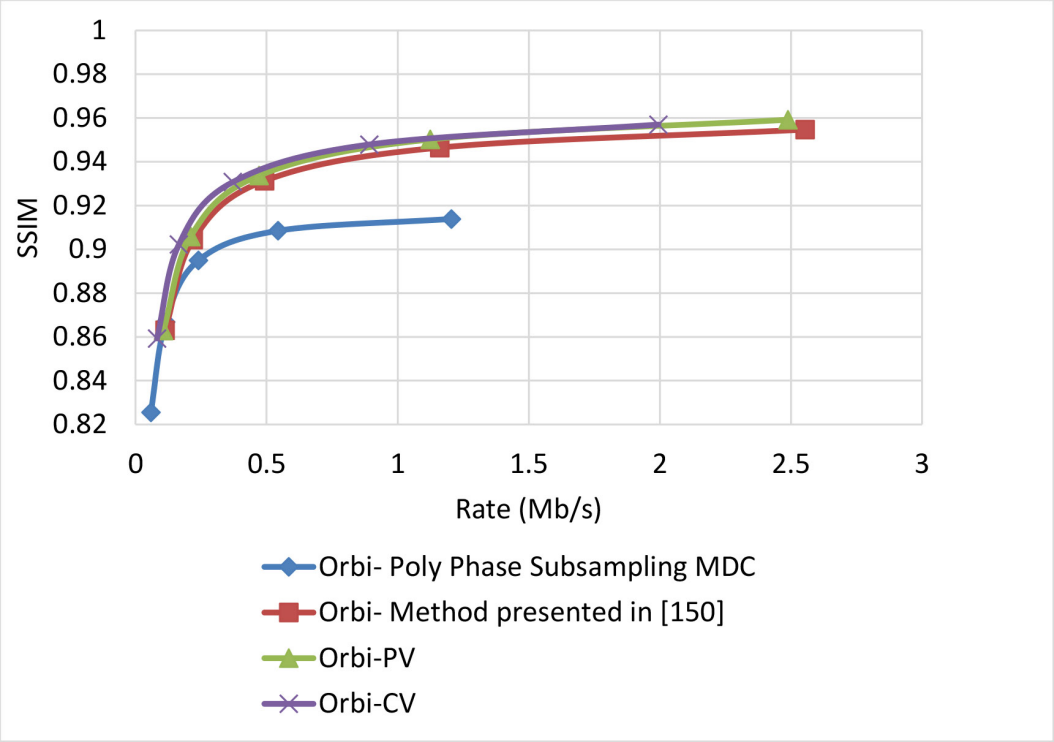


Figure 4.6: Average SSIM assessment of video "Orbi" (color image).

When it comes to the evaluation of the proposed algorithm for the reconstructed depth map image, it shows a better performance. As shown in Figure 4.7 and Figure 4.8 for the video Interview and in Figure 4.9 and Figure 4.10 for the video Orbi, the improvement of the proposed algorithm is considerably evident. This can be due to the fact that metrics PV and CV are calculated based on the depth map image and therefore blocks with larger values of metrics PV and CV can be considered as the least predictable blocks in the depth map image. Therefore, focusing on these pixels in each description results in a more accurate reconstruction in the decoder. In view of the PSNR assessment, about 8 dB for video Interview and more than 10 dB for video Orbi improvement have been achieved by the proposed algorithm. Such high performance of the proposed algorithm in view of the SSIM assessment is also more evident compared with the color video assessment. With regards to the SSIM assessment, the proposed algorithm outperforms by more than 0.02 compared to the PSS-MDC method.

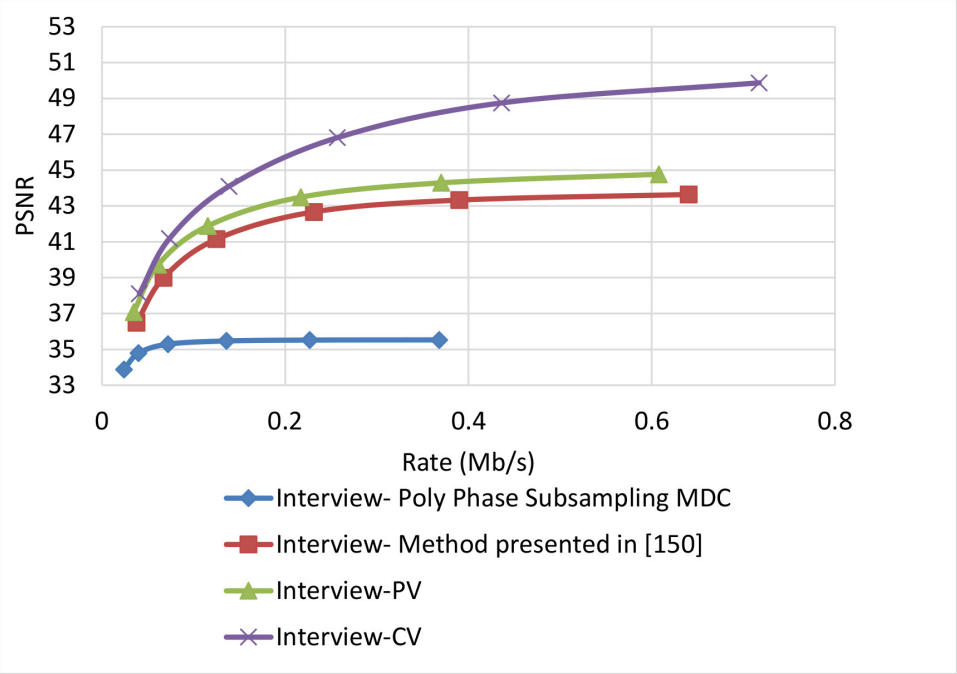


Figure 4.7: Average PSNR assessment of video "Interview" (depth image).

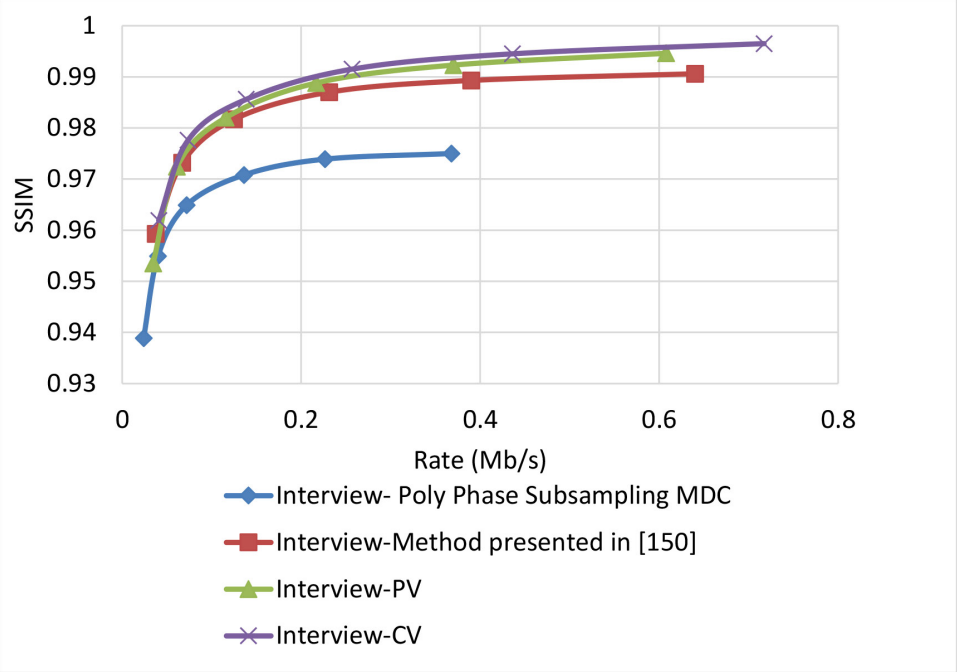


Figure 4.8: Average SSIM assessment of video "Interview" (depth image).

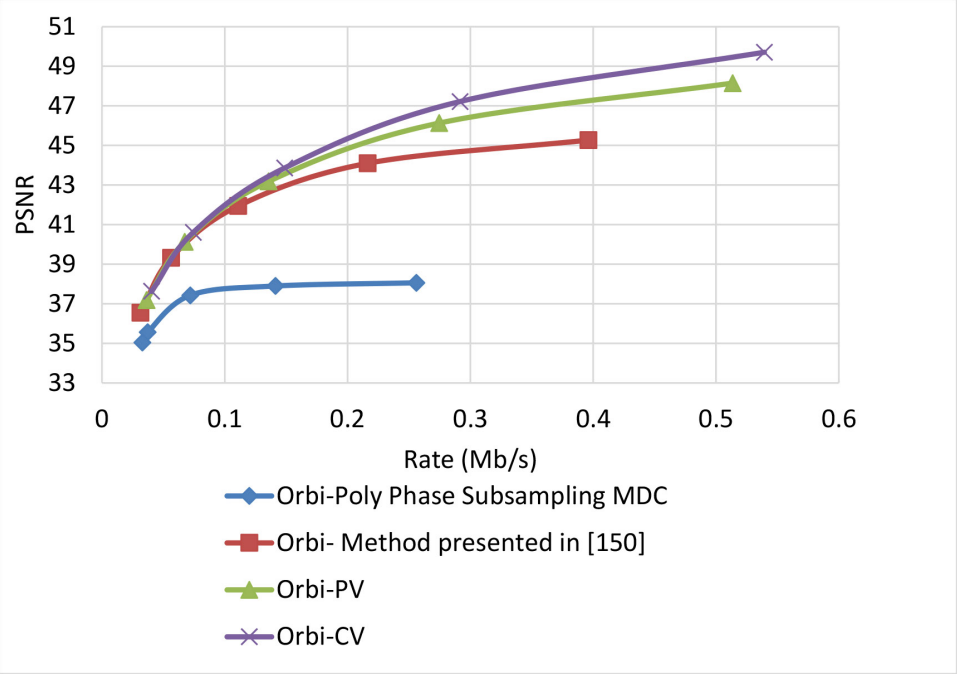


Figure 4.9: Average PSNR assessment of video "Orbi" (depth image).

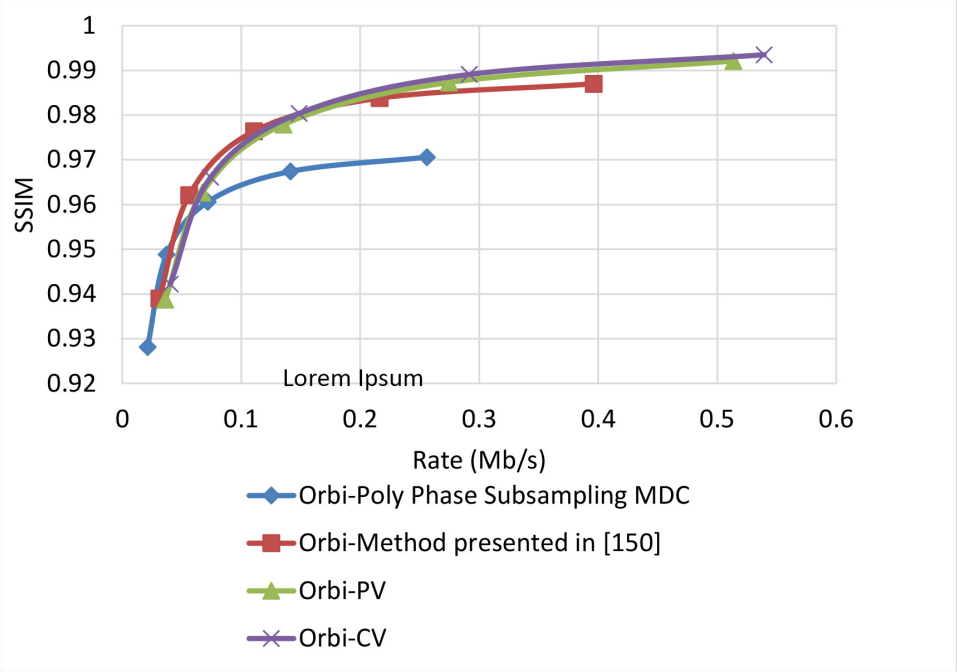


Figure 4.10: Average SSIM assessment of video "Orbi" (depth image).

Figures 4.11 and 4.12 show the PSNR assessment for two different test video sequences, called "Ballet" and "Breakdancers", generated by the interactive visual media group at Microsoft research [149]. Unlike the previous test video sequences, the new test video sequences include objects with very fast movement. As can be seen in these figures, like previous experiments the proposed MDC method provide improved performance. Figures 4.13 and 4.14 demonstrate the PSNR assessment of these two test video sequences using the most recent video encoder, i.e. H.265/HEVC. To implement H.265/HEVC encoder, we used H.265 reference software, HM 6.0 [147].

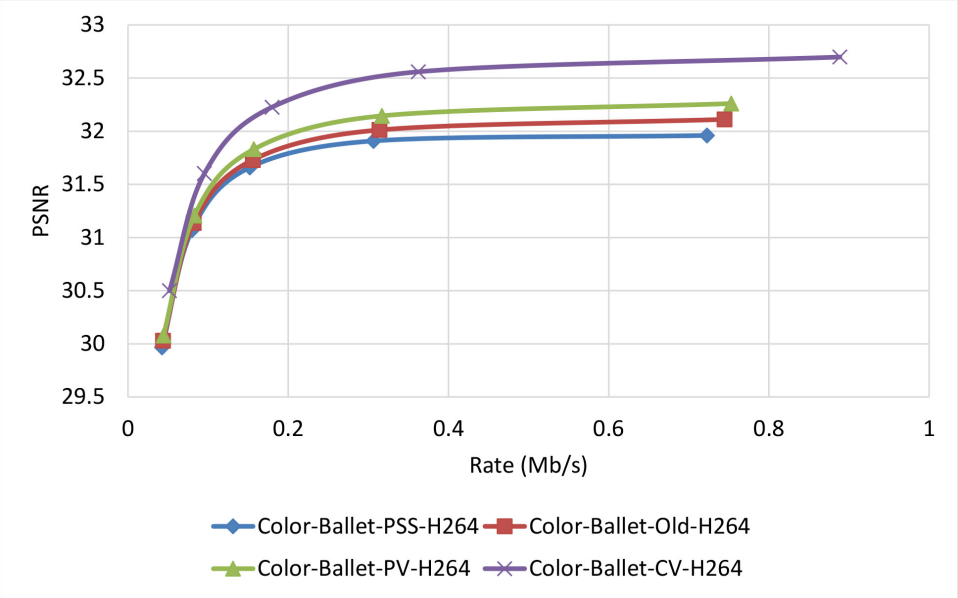


Figure 4.11: Average PSNR assessment of video Ballet using H.264 encoder (color Image).

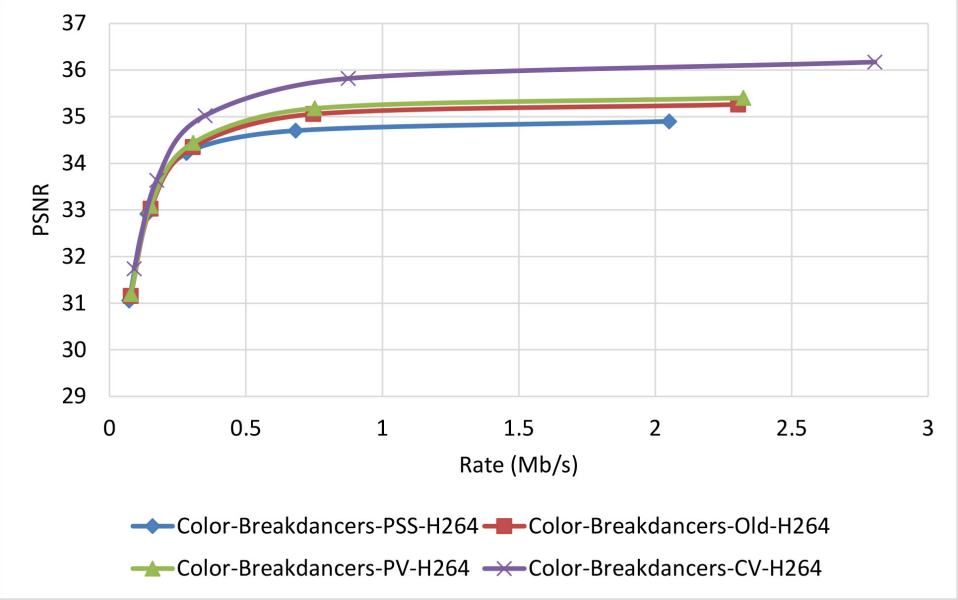


Figure 4.12: Average PSNR assessment of video Breakdancers using H.264 encoder (color Image).

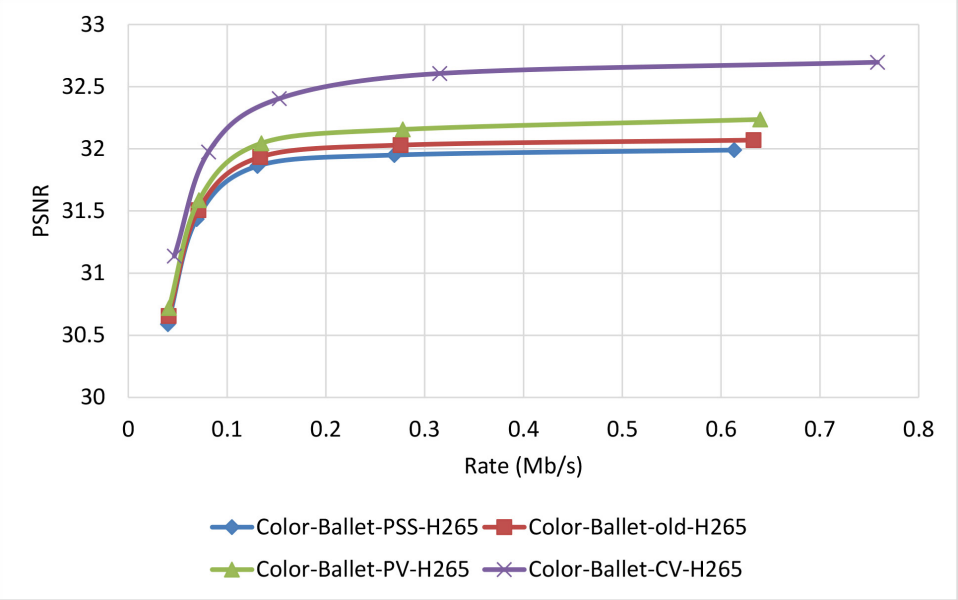


Figure 4.13: Average PSNR assessment of video Ballet using H.265 encoder.

Before evaluating the hybrid MDC method, we compare the temporal MDC method and primary spatial MDC method performance. Figure 4.15 shows the PSNR assessment of both MDC types with two or four descriptions. In this test, the colour image is partitioned spatially and temporally into two or four descriptions

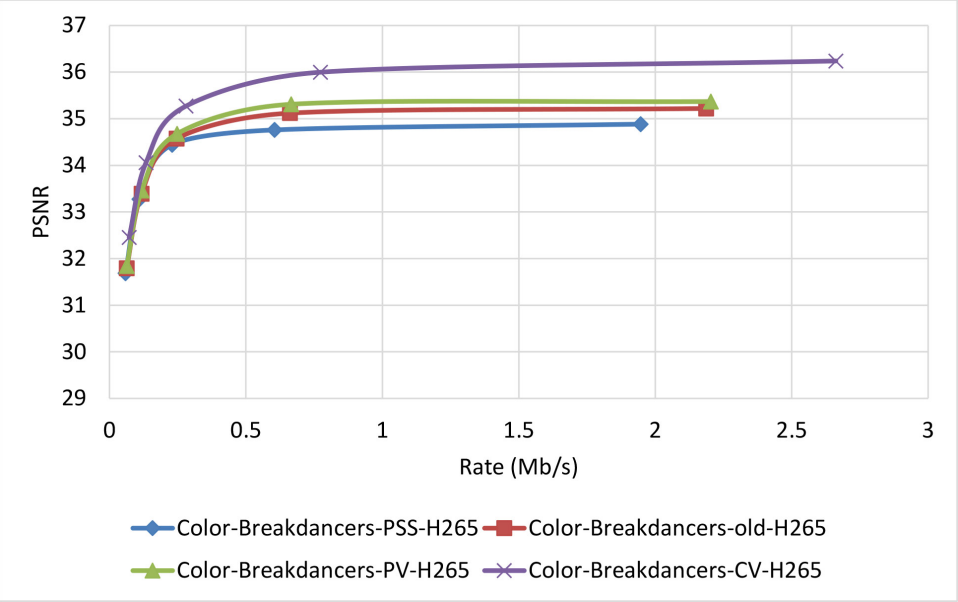


Figure 4.14: Average PSNR assessment of video Breakdancers using H.265 encoder.

and streamed toward the receiver. In the decoder, it is assumed that only one description is available and others are missed. Then the missed information is estimated from the available description. As shown in Figure 4.15, temporal MDC performs much better than spatial MDC in point of PSNR assessment, however, spatial MDC is more robust against noise variation compared to the temporal MDC. As shown in Figure 4.15 the slope of the graphs related to the spatial MDC for the large rate is approximately zero while the slope for the temporal MDC is not zero. That means that temporal MDC is more sensitive to noise compared to spatial MDC.

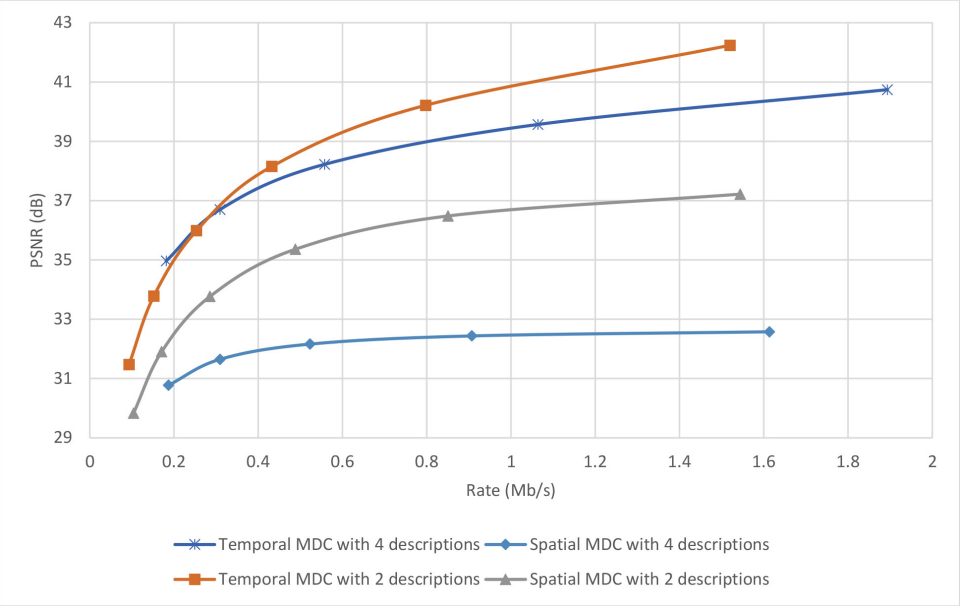


Figure 4.15: Temporal MDC vs spatial MDC performance comparison for the video "Interview".

Also, it should be noted that these results are the average PSNR assessment of all 90 frames. In fact, the successfully received frames are decoded with a PSNR assessment around 54 dB for the large rate, while the missed frames are decoded with PSNR about 37 dB. So there is a huge distance between the frame that received successfully and the one estimated. For the spatial MDC, all frames are decoded with PSNR assessment approximately 38 dB. It is worth mentioning that the test video sequence selected for this test has very low movement and frames are very dependent; therefore, the temporal MDC provides much better results.

Figure 4.16 and Figure 4.17 show the PSNR assessment of the proposed MDC method with spatiotemporal enhancement algorithm for the color image of video "Interview" and "Orbi", respectively. The graphs compare the performance of the proposed method and previous methods. As can be seen in Figure 4.16, the quality of the reconstructed video of the test video sequence "Interview" is improved by about 6 to 7 dB. First of all, this huge gap is due to the very low movement of the background that we have in this test video sequence; therefore adding temporal information improves the PSNR assessment significantly. As argued before, the slope of PSNR assessment for the proposed method is smaller than the slope of the temporal MDC and greater than the slope of the spatial MDC presented in Figure 4.15; so

it is less sensitive to the noise variation compared to a pure temporal MDC type. More importantly, the proposed method provides better performance compared to the spatial MDC type. In Figure 4.17, we observe similar results for the test video "Orbi", however, the improvement is not as large as the improvement achieved for the test video "Interview". As shown in Figure 4.17, the proposed method outperforms about 2 dB compared to previous methods.

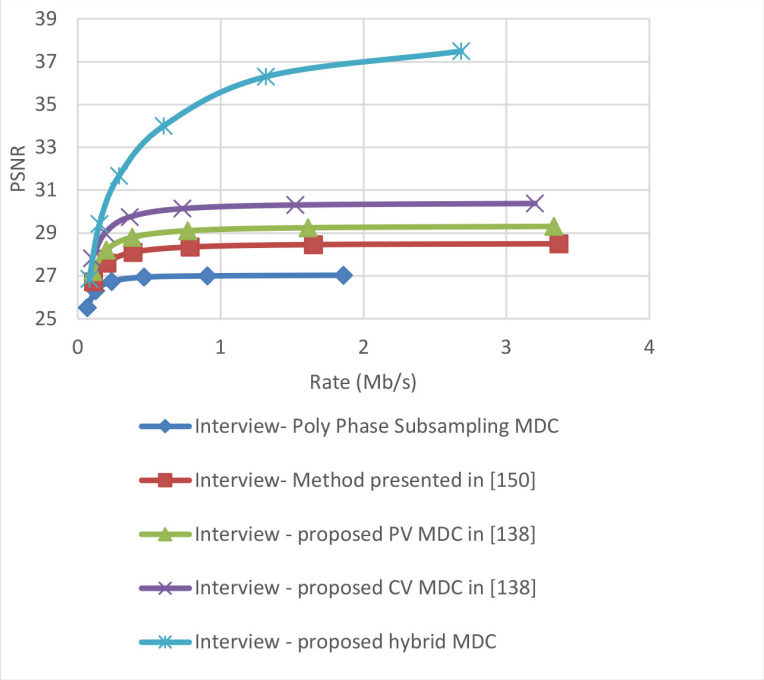


Figure 4.16: Average PSNR assessment video "interview" usig hybrid MDC (color image).

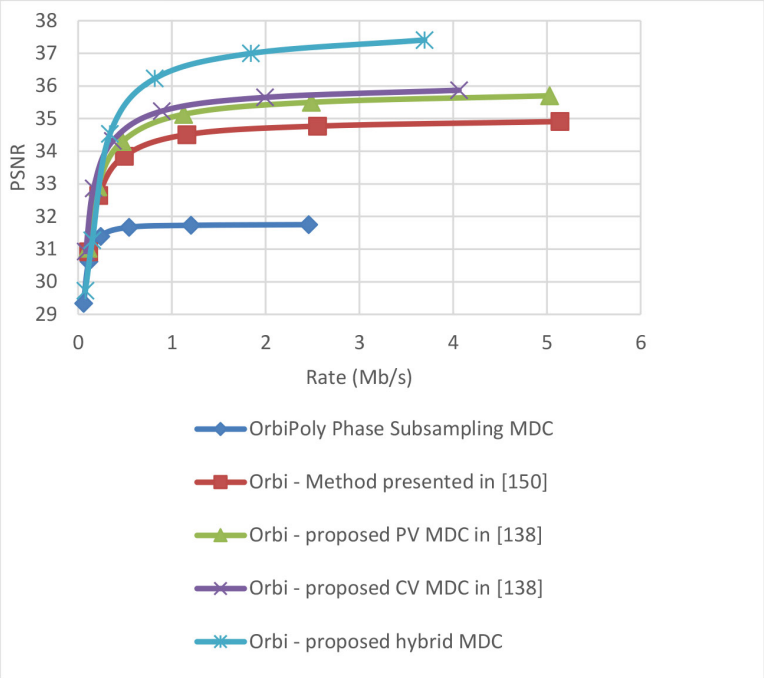


Figure 4.17: Average PSNR assessment video "Orbi" usig hybrid MDC (color image).

4.3.2 Test Results of Scenario Two

In Section 4.3.1, we assumed only one description of four descriptions is available, which would result in a loss of 75% of the content, although it is not a realistic scenario. As a way to simulate a real-world scenario, we examined our method with a random packet loss scenario. Congestion on routers causes packets to be dropped as a result of queues being full, which lead to packet loss. There are also instances when links or network devices contain errors, causing packets to be incompletely delivered. Other link indicators can be used to assess network quality; however, we evaluate a link's quality primarily based on the packet losses.

Quality of service is closely related to packet loss. A packet loss rate that is acceptable for one application might not be acceptable for another. Depending on the data being sent, the packet loss varies. Generally, packet loss between 5% and 10% of the total packet stream will significantly affect quality.

In audio and video applications specifically, packet loss less than 1% is considered good [153]. The link quality is marked as "acceptable" if it is between 1% and 2.5%. We have poor network quality if packet loss is around 2.5%-5%. 5%-12% packet

loss consider if the network is very poor. If packet loss is greater than 12%, it is bad. Packet loss above 10%-12% is unacceptable. In general, above 10%-12% packet loss, there is an unacceptable level of packet losses, causing highly long timeout connections, and video conferencing is unusable [154].

For this scenario we are using the video "pingpong" which include 150 frames (5 sec) captured by a 2 HD cameras. We segmented the video to 37 segments each with 4 frames and encode with JM software, with I frames are repeated every 4 frames and only P frames are used between I frames. The result of having a 2.5% or 5% random packet loss on the streamed video is shown in Table 4.4. we simulated our proposed method and compared the result with the poly phase subsampling MDC method. As can be seen around 0.5 dB improvement is achieved when the 2.5% packet loss applied and around 1 db improvement is seen when when packet loss rate is 5%.

Table 4.4: Performance assessment of the proposed method with packet drop for $Q = 30$.

	Packet drop rate		Number of packets per segment
	2.5%	5%	
Average PSNR for primary spatial MDC with 4 descriptions	33.59	30.35	8
Average PSNR for proposed MDC method with spatial enhancement algorithm	34.01	31.15	10

4.4 Conclusion

Throughout this chapter, we presented our simulation results based on our MDC algorithm evaluation. Two test scenarios were conducted. In the first scenario, it was assumed only three descriptions were lost at receiver, and the decoder estimated the lost descriptions from the only available description. Based on the objects present in a scene, our algorithm allocates more bandwidth to the region of interest. Human eyes are more sensitive to objects than pixels, so the proposed algorithm can provide a better performance than the PSS MDC in terms of subjective assessment. Nevertheless, the objective assessment results confirm that the proposed spatial MDC algorithm has improved performance. The proposed algorithm enhances the current basic decimation to a non-identical decimation for depth map images. The second test scenario used a random packet failure rate of 2.5% or 5% to simulate poor network quality and determine if our method is applicable in real life.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Multimedia applications are growing increasingly common due to digital communication developments over the past few decade. Nowadays, TV on-demand applications such as Netflix, Amazon, Disney, etc are using available bandwidth more than other types of applications, and operators are trying to increase their capacities to be able to support their customers. In addition to the limited capacity, multimedia streaming is affected by packet failure in the network due to packet loss, packet corruption, and large packet delay. An appropriate solution against packet failure in the error-prone environment can be multiple description coding (MDC). With MDC, one video description is partitioned into several separately decodable descriptions. In the instance of missing a description during transmission, the decoder is capable to estimate the lost description from other error-free description(s).

Live sport streaming events as one of popular video broadcast services are attracting large number of audiences, nowadays. This type of application requires low delay, long distance, and high quality transmission. Such requirements suggest to avoid temporal MDC because of increased delay or frequency MDC due to complexity and lower performance. To improve the primary spatial partitioning (applicable to 3D videos), a nonidentical decimation algorithm for the stereoscopic videos has been provided in this thesis. Our algorithm works based on existing objects in the scene and assigns more bandwidth to the region of interest. The assessment results confirm the improved performance achieved by the proposed spatial MDC algorithm.

With regard to the depth map image, most parts of the depth map have similar depth values and therefore decimation in those parts can save bandwidth or storage

without considerable quality degradation. However, for the parts of the frame with high pixels' value variation, it is recommended to keep the original resolution. Therefore, with the new MDC methods, those parts of the depth map image that have large variations are encoded with the original resolution.

5.2 Future Work

What we did in this thesis are, first, detecting Region of Interest, and second, creating MDC descriptions while assigning more bandwidth to the Region of Interest. Our method allocated more bandwidth to the RoI via increasing its spatial resolution. In the second introduced MDC method, i.e, spatiotemporal MDC, we decreased the bandwidth assignment to the background to reduce the total bandwidth assignment (that had been increased by the first introduced MDC method). Because the background region can be estimated from the previous frames with a better quality compared to spatial estimation. In this section, we discuss a number of interesting possibilities for the extension of the current work:

- Modify the description enhancement algorithm to enhance the quality of different regions using Q parameter instead of the spatial resolution. This way we will have a more precise adjustment tools to increase or decrease the bandwidth assignment to different regions. With the current method we only able to double or quadruple the bandwidth assignment to RoI. Using Q parameter as the adjustment tools, we will able to increment the bandwidth assignment more precisely. In addition to such flexible adjustment tool, we will able to have different number of descriptions as well.

- This method can be combined with the layered coding to provide flexibility to the video stream. The description enhancement algorithm can add the enhancement information to a enhancement layer and not to the base layer (which is done in the current method). This way encoder streams the enhancement layers if there is still available bandwidth after streaming of the base layer.

List of References

[1] J. Patterson, "A history of 3D cinema." *The Gaurdian*, https://www.theguardian.com/film/2009/aug/20/3d-film-history, Accessed: 1-Dec-2021.

[2] C. Hewage, *3D Video Processing and Transmission Fundamentals*. bookboon.com, 2014.

[3] L. Meesters, W. IJsselsteijn, and P. Seuntiens, "A survey of perceptual evaluations and requirements of three-dimensional TV," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 381–391, March 2004.

[4] J. Clement, "Consumer internet video traffic in north america from 2016 to 20211." *Statista*, https://www.statista.com/statistics/267213/forecast-for-the-data-volume-of-internet-video-communications/, Accessed: 1-Dec-2021.

[5] J. Flint and M. Maidenberg, "Netflix adds 16 million new subscribers as homebound consumers stream away," *Wall St. J. (East Ed)*, Apr. 2020.

[6] V. Fosty and T. Houben, "Online media streaming will benefit from the coronavirus pandemic." *Deloitte.com*, https://www2.deloitte.com/be/en/pages/technology-media-and-telecommunications/articles/online-media-streaming.html, Accessed: 1-May-2021.

[7] "Cisco annual internet report (20182023)." *Cisco*, https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html, Accessed: 24-Dec-2021.

[8] T. Dreier, "On-demand viewing growing much faster than live, says conviva." *Streamingmedia.com*, https://www.streamingmedia.com/Articles/News/Online-Video-News/On-Demand-Viewing-Growing-Much-Faster-Than-Live-Says-Conviva-133418.aspx, Accessed: 24-Dec-2021.

[9] "47 must-know live video streaming statistics." https://livestream.com/blog/62-must-know-stats-live-video-streaming, Accessed: 24-Dec-2021.

[10] C. Gough, "Sports on U.S. TV-statistics and facts." *Statista*, https://www.statista.com/topics/2113/sports-on-tv/#dossierKeyfigures, Accessed: 24-Dec-2021.

[11] "FIFA financial report 2018." *FIFA*, https://digitalhub.fifa.com/m/337fab75839abc76/original/xzshsoe2ayttyquuxhq0-pdf.pdf, Accessed: 24-Dec-2021.

[12] "Current world population." *Worldometers*, https://www.worldometers.info/world-population/, Accessed: 29-April-2021.

[13] K. Calagari, M. Elgharib, S. Shirmohammadi, and M. Hefeeda, "Sports VR content generation from regular camera feeds," in *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, (New York, NY, USA), p. 699707, Association for Computing Machinery, 2017.

[14] A. Smolic and H. Kimata, "Report on status of 3DAV exploration," Tech. Rep. N5558, ISO/IEC JTC1/SC29/WG11, Thailand, March 2003.

[15] M. Wien, *High Efficiency Video Coding (HEVC): Coding Tools and Specification*. Springer International Publishing, 2014.

[16] ITU, "Stereoscopic television MPEG-2 multi-view profile," Tech. Rep. Report BT.2017-0, Itu, 1998.

[17] A. Smolic and H. Kimata, "Report on status of 3DAV exploration," Tech. Rep. W5877, ISO/IEC JTC1/SC29/WG11, Norway, July 2003.

[18] "Applications and requirements for 3DAV," Tech. Rep. N5877, ISO/IEC JTC1/SC29/WG11, Norway, July 2003.

[19] P. Merkle, K. Muller, A. Smolic, and T. Wiegand, "Efficient compression of multi-view video exploiting inter-view dependencies based on H.264/MPEG4-AVC," in *2006 IEEE International Conference on Multimedia and Expo*, pp. 1717–1720, 2006.

[20] "One-way transmission time." ITU Recommendation G.114 (05/03), https://www.itu.int/rec/T-REC-G.114, Accessed: 24-Dec-2021.

[21] "Defining broadband: Network latency and application performance." *ADTRAN*, https://ecfsapi.fcc.gov/file/6520222942.pdf, Accessed: 24-Dec-2021.

[22] "Measuring broadband canada." *Government of Canada, Canadian Radiotelevision and Telecommunications Commission (CRTC), https://crtc.gc.ca/eng/publications/reports/rp200601/rp200601.htm, Accessed: 24-Dec-2021*,.

[23] "End-user multimedia QoS categoriese." *ITU Recommendation G.1010*, https://www.itu.int/rec/T-REC-G.1010-200111-I, Accessed: 24-Dec-2021.

[24] "3rd generation partnership project; technical specification group services and system aspects; service aspects; services and service capabilities (release 9)." *3*GPP TS 22.105 V9.0.0, https://www.arib.or.jp/english/html/overview/doc/STD-T63v9_60/5_Appendix/Rel9/21/21902-900.pdf, Accessed: 24-Dec-2021.

[25] R. F. Tim Rahrer and S. Wright, "Triple-play services quality of experience (QoE) requirements." *Broadband Forum*, Dec 2006, `https://www.broadband-forum.org/download/TR-126.pdf`, Accessed: 24-Dec-2021.

[26] M. Kazemi, *Multiple description video coding based on base and enhancement layers of SVC and channel adaptive optimization*. PhD thesis, Sharif University of Technology, Tehran, Iran, 2012.

[27] M. Kazemi, S. Shirmohammadi, and K. H. Sadeghi, "A review of multiple description coding techniques for error-resilient video delivery," *Multimedia Systems*, vol. 20, p. 283309, June 2014.

[28] W. Welling, *Photography in America: The Formative Years,1839-1900*. Crowell New York, 1978.

[29] I. P. Howard and B. J. Rogers, *Binocular Vision and Stereopsis*. New York: Oxford University, 1995.

[30] S. S. Bhattacharyya, F. Deprettere, R. Leuperse, and J. Takala, *Handbook of Signal Processing Systems*. Springer International Publishing, 2010.

[31] M. Tanimoto, M. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint tv," *IEEE Signal Processing Magazine*, vol. 28, pp. 67–76, Jan 2011.

[32] E. Boult, R. Micheals, M. Eckmann, C. P. X. Gao, and S. Sablak, "Omnidirectional video applications," in *Proceedings of the 8'h International Symposium on Intelligent Robotic Systems*.

[33] E. Kurutepe, M. Civanlar, and A. Tekalp, "Client-driven selective streaming of multiview video for interactive 3DTV," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 1558–1565, Nov 2007.

[34] M. Pollefeys, R. Koch, M. Vergauwen, and L. Van Gool, "Hand-held acquisition of 3D models with a video camera," in *Second International Conference on 3-D Digital Imaging and Modeling*, pp. 14–23, 1999.

[35] O. Wilinski and K. V. Overveld, "Depth from motion using confidence based block matching," in *Proceedings of Image and Multidimensional Signal Processing Workshop*, pp. 159– 192, July 1998.

[36] P. Harman, J. Flack, S. Fox, and M. Dowley, "Rapid 2D to 3D conversion," *SPIE: Stereoscopic Displays and Virtual Reality Systems*, vol. 4660, pp. 78– 86, 2002.

[37] Q. Deng, Y. Zhang, and S. Li, "The overview of 2D to 3D conversion," in *2013 IEEE 4th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pp. 226–229, Nov 2013.

[38] X. Cao, A. Bovik, Y. Wang, and Q. Dai, "Converting 2D video to 3D: An efficient path to a 3D experience," *IEEE MultiMedia*, vol. 18, pp. 12–17, April 2011.

[39] Y.-L. Chang, Y.-P. Tsai, T.-H. Chang, Y.-R. Chen, and S. Lei, "A depth map refinement algorithm for 2D-to-3D conversion," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1437–1440, March 2012.

[40] J. Herrera, C. del Blanco, and N. Garca, "Edge-based depth gradient refinement for 2D to 3D learned prior conversion," in *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2015*, pp. 1–4, July 2015.

[41] Y.-K. Lai, Y.-F. Lai, and Y.-C. Chen, "An effective hybrid depth-generation algorithm for 2D-to-3D conversion in 3D displays," *Journal of Display Technology*, vol. 9, pp. 154–161, March 2013.

[42] L. Sisi, W. Fei, and L. Wei, "The overview of 2D to 3D conversion system," in *2010 IEEE 11th International Conference on Computer-Aided Industrial Design Conceptual Design (CAIDCD)*, vol. 2, pp. 1388–1392, Nov 2010.

[43] J. Konrad, M. Wang, P. Ishwar, C. Wu, and D. Mukherjee, "Learning-based, automatic 2D-to-3D image and video conversion," *IEEE Transactions on Image Processing*, vol. 22, pp. 3485–3496, Sept 2013.

[44] A. Box tech. rep., 2014.

[45] A. Woods, T. Docherty, and R. Koch, "Image distortions in stereoscopic video systems," *SPIE: Stereoscopic Displays and Virtual Reality Systems*, vol. 1915, pp. 36–48, Feb. 1993.

[46] V. V. Petrov and K. A. Grebenyuk, "Optical correction of depth plane curvature image distortion," *Proc. SPIE*, vol. 6637, pp. 66370P–66370P–6, 2007.

[47] G. Iddan and G. Yahav, "3D imaging in the studio (and elsewhere ...)," *SPIE: Stereoscopic Displays and Virtual Reality Systems*, vol. 4298, pp. 48– 55, 2001.

[48] M. Kawakita, K. Iizuka, T. Aida, H. Kikuchi, H. Fujikake, J. Yonai, and K. Takizawa, "Axi-vision camera (real-time distance-mapping camera)," *Appl. Opt.*, vol. 39, pp. 3931–3939, Aug 2000.

[49] J. Choi, *Range Sensors: Ultrasonic Sensors, Kinect, and LiDAR*, pp. 2521–2538. Dordrecht: Springer Netherlands, 2019.

[50] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," *SPIE: Stereoscopic Displays and Virtual Reality Systems*, vol. 5291, pp. 93– 104, 2004.

[51] S. Chabatkova, "IPTV and video statistics." `https://www.datasciencecentral.com/profiles/blogs/iptv-and-video-statistics`, Accessed: 26-Dec-2021.

[52] H. Hai, "Automatic repeat-request courseware." Vaasa University of Applied Sciences, `https://www.theseus.fi/bitstream/handle/10024/56440/Automatic%20Repeat-Request%20Courseware.pdf?sequence=1&isAllowed=y`, Accessed: 29-April-2021.

[53] Y. Alotaibi, "A new multi-path forward error correction (fec) control scheme with path interleaving for video streaming," in *2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 1655–1660, 2015.

[54] B. Zhang, P. Cosman, and L. B. Milstein, "Energy optimization for wireless video transmission employing hybrid arq," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 5606–5617, 2019.

[55] H. S. Oh, J. G. Shin, W. S. Jeon, and D. G. Jeong, "Reliable video multicast based on al-fec and h.264/avc video traffic characteristics," in *2017 14th IEEE Annual Consumer Communications Networking Conference (CCNC)*, pp. 945–950, 2017.

[56] J. Wu, R. Tan, and M. Wang, "Streaming high-definition real-time video to mobile devices with partially reliable transfer," *IEEE Transactions on Mobile Computing*, vol. 18, no. 2, pp. 458–472, 2019.

[57] M.-T. Sun and A. R. Reibman, *Compressed Video Over Networks (Signal Processing and Communications)*. CRC Press, September 2000.

[58] P. Lambert, W. De Neve, Y. Dhondt, and R. Van de Walle, "Flexible macroblock ordering in h.264/avc," *Journal of Visual Communication and Image Representation*, vol. 17, no. 2, pp. 358–375, 2006. Introduction: Special Issue on emerging H.264/AVC video coding standard.

[59] J. Wang, S. Li, K. Shimizu, T. Ikenaga, and S. Goto, "Unequal error protected transmission with dynamic classification in h.264/avc," in *2007 7th International Conference on ASIC*, pp. 798–801, 2007.

[60] J. Nightingale, Q. Wang, C. Grecos, and S. Goma, "Video adaptation for consumer devices: opportunities and challenges offered by new standards," *IEEE Communications Magazine*, vol. 52, pp. 157–163, December 2014.

[61] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the h.264/avc standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 1103–1120, Sept 2007.

[62] Y.-C. Lee, J. Kim, Y. Altunbasak, and R. M. Mersereau, "Layered coded vs. multiple description coded video over error-prone networks," *Signal Processing: Image Communication*, vol. 18, no. 5, pp. 337 – 356, 2003.

[63] J. Chakareski, S. Han, and B. Girod, "Layered coding vs. multiple descriptions for video streaming over multiple paths," *Multimedia Systems*, vol. 10, no. 4, pp. 275–285, 2005.

[64] A. Reibman, Y. Wang, X. Qiu, Z. Jiang, and K. Chawla, "Transmission of multiple description and layered video over an egprs wireless network," in *2000 International Conference on Image Processing*, vol. 2, pp. 136–139 vol.2, Sept 2000.

[65] R. Singh, A. Ortega, L. Perret, and W. Jiang, "Comparison of multiple description coding and layered coding based on network simulations," 2000.

[66] "High efficiency video coding (HEVC) scalable extension draft 4." *JCT-VC*, JCTVC-O1008-v3.

[67] S. Radhakrishnan, *Effective video coding for multi media application.* InTech, 2011.

[68] J. Boyce, Y. Ye, J. Chen, and A. Ramasubramonian, "Overview of SHVC: Scalable extensions of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2015.

[69] T. Wiegand, G. J. Sullivan, J. Reichel, H. Schwarz, and M.Wien, "Joint draft itu-t rec. h.264iiso!iec 14496-10 i amd.3 scalable video coding," Tech. Rep. JVT-X201, ISO/IEC JTC1/SC29/WG11 and ITU SG16 Q.6, Geneva, Switzerland, July 2007.

[70] Z. Liu, G. Cheung, J. Chakareski, and Y. Ji, "Multiple description coding and recovery of free viewpoint video for wireless multi-path streaming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, pp. 151–164, Feb 2015.

[71] S. Shirani, M. Gallant, and F. Kossentini, "Multiple description image coding using pre- and post-processing," in *International Conference on Information Technology: Coding and Computing*, pp. 35–39, Apr 2001.

[72] M. Gallant, S. Shirani, and F. Kossentini, "Standard-compliant multiple description video coding," in *2001 International Conference on Image Processing*, vol. 1, pp. 946–949 vol.1, 2001.

[73] T. Tillo and G. Olmo, "Data-dependent pre- and postprocessing multiple description coding of images," *IEEE Transactions on Image Processing*, vol. 16, pp. 1269–1280, May 2007.

[74] Y. Yapc, B. Demir, S. Ertrk, and O. Urhan, "Down-sampling based multiple description image coding using optimal filtering," *SPIE: journal of Electronic Imaging*, vol. 17, 2008.

[75] C. Ates, Y. Urgun, B. Demir, O. Urhan, and S. Erturk, "Polyphase down-sampling based multiple description image coding using optimal filtering with flexible redundancy insertion," in *International Conference on Signals and Electronic Systems*, pp. 193–196, Sept 2008.

[76] S. Shirani, "Content-based multiple description image coding," *IEEE Transactions on Multimedia*, vol. 8, pp. 411–419, April 2006.

[77] J. Wang and J. Liang, "H.264 intra frame coding and JPEG 2000-based predictive multiple description image coding," in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pp. 569–572, Aug 2007.

[78] Z. Wei, K.-K. Ma, and C. Cai, "Prediction-compensated polyphase multiple description image coding with adaptive redundancy control," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, pp. 465–478, March 2012.

[79] S. Zhu, Z. He, X. Meng, J. Zhou, Y. Guo, and B. Zeng, "A new polyphase down-sampling-based multiple description image coding," *IEEE Transactions on Image Processing*, vol. 29, pp. 5596–5611, 2020.

[80] W. Jiang and A. Ortega, "Multiple description coding via polyphase transform and selective quantization," *SPIE: journal of Electronic Imaging*, vol. 3653, pp. 998–1008, 1999.

[81] C. Zhu and M. Liu, "Multiple description video coding based on hierarchical b pictures," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, pp. 511–521, April 2009.

[82] A. Zhao, W. Wang, H. Cui, and K. Tang, "Efficient multiple description scalable video coding scheme based on weighted signal combinations," *Tsinghua Science and Technology*, vol. 12, pp. 86–90, Feb 2007.

[83] D. T. Duong, "New h.266/vvc based multiple description coding for robust video transmission over error-prone networks," in *2020 International Conference on Advanced Technologies for Communications (ATC)*, pp. 155–159, 2020.

[84] B. Bross, J. Chen, S. Liu, and Y.-K. Wang, "doc. jvet-q2001 of itu-t/iso/iec joint video exploration team (jvet)," in *Versatile Video Coding (Draft 8)*, vol. 17th meeting, 2020.

[85] ITU, "Beyond HEVC: Versatile video coding project starts strongly in joint video experts team." `https://news.itu.int/versatile-video-coding-project-starts-strongly/`, Apr. 2018. Accessed: 2021-6-23.

[86] S. Milani and G. Calvagno, "Multiple description distributed video coding using redundant slices and lossy syndromes," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 51–54, 2010.

[87] J. Apostolopoulos, "Reliable video communication over lossy packet networks using multiple state encoding and path diversity," *Visual Communications and Image Processing (VCIP)*, 2001.

[88] G. A. Thomas, "Television motion measurement for datv and other applications," Tech. Rep. BBC RD 1987/11, Research Department, Engineering Division The Britich Broadcasting Corporation, September 1987.

[89] M. Zhang, W. Liu, R. Wang, and H. Bai, "A novel multiple description video coding algorithm," in *International Conference on Computational Intelligence and Security*, vol. 1, pp. 550–553, Dec 2008.

[90] H. Bai, Y. Zhao, and C. Zhu, "Multiple description video coding using adaptive temporal sub-sampling," in *2007 IEEE International Conference onMultimedia and Expo*, pp. 1331–1334, July 2007.

[91] R. Kibria and J. Kim, "H.264/avc-based multiple description coding for wireless video transmission," in *Proceedings of the 12th WSEAS International Conference on Communications*, ICCOM'08, (Stevens Point, Wisconsin, USA),

pp. 429–432, World Scientific and Engineering Academy and Society (WSEAS), 2008.

[92] T. Tillo and G. Olmo, "Low complexity pre postprocessing multiple description coding for video streaming," in *2004 International Conference on Information and Communication Technologies: From Theory to Applications, 2004. Proceedings*, pp. 519–520, April 2004.

[93] T. Tillo, E. Baccaglini, and G. Olmo, "Multiple descriptions based on multirate coding for jpeg 2000 and h.264/avc," *IEEE Transactions on Image Processing*, vol. 19, pp. 1756–1767, July 2010.

[94] I. Radulovic, P. Frossard, Y.-K. Wang, M. Hannuksela, and A. Hallapuro, "Multiple description video coding with h.264/avc redundant pictures," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, pp. 144–148, Jan 2010.

[95] V. Parameswaran, A. Kannur, and B. Li, "Adapting quantization offset in multiple description coding for error resilient video transmission," *Journal of Visual Communication and Image Representation*, vol. 20, no. 7, pp. 491 – 503, 2009.

[96] V. Vaishampayan, "Design of multiple description scalar quantizers," *IEEE Transactions on Information Theory*, vol. 39, pp. 821–834, May 1993.

[97] C. Tian and S. Hemami, "A special class of multiple description scalar quantizers," in *IEEE Information Theory Workshop*, pp. 135–140, Oct 2004.

[98] A. Reibman, H. Jafarkhani, Y. Wang, and M. Orchard, "Multiple description video using rate-distortion splitting," in *2001 International Conference on Image Processing, 2001. Proceedings*, vol. 1, pp. 978–981 vol.1, 2001.

[99] K. Matty and L. Kondi, "Balanced multiple description video coding using optimal partitioning of the dct coefficients," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 928–934, July 2005.

[100] D. Comas, R. Singh, A. Ortega, and F. Marques, "Unbalanced multiple-description video coding with rate-distortion optimization," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 1, p. 418296, 2003.

[101] D.-M. Chung and Y. Wang, "Multiple description image coding using signal decomposition and reconstruction based on lapped orthogonal transforms," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, pp. 895–908, Sep 1999.

[102] G. Sun, U. Samarawickrama, J. Liang, C. Tian, C. Tu, and T. Tran, "Multiple description coding with prediction compensation," *IEEE Transactions on Image Processing*, vol. 18, pp. 1037–1047, May 2009.

[103] V. Goyal, J. Kovacevic, R. Arean, and M. Vetterli, "Multiple description transform coding of images," in *1998 International Conference on Image Processing*, vol. 1, pp. 674–678 vol.1, Oct 1998.

[104] Y. Wang, M. Orchard, and A. Reibman, "Optimal pairwise correlating transforms for multiple description coding," in *1998 International Conference on Image Processing,ICIP 98*, vol. 1, pp. 679–683 vol.1, Oct 1998.

[105] Y. Wang, M. Orchard, V. Vaishampayan, and A. Reibman, "Multiple description coding using pairwise correlating transforms," *IEEE Transactions on Image Processing*, vol. 10, pp. 351–366, Mar 2001.

[106] Y. Wang, A. Reibman, M. Orchard, and H. Jafarkhani, "An improvement to multiple description transform coding," *IEEE Transactions on Signal Processing*, vol. 50, pp. 2843–2854, Nov 2002.

[107] "Video codec for audiovisual services at p64 kbit/s," *ITU-T Rec. H.261*, 1988.

[108] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h.264/avc video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 560–576, July 2003.

[109] G. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, pp. 1649–1668, Dec 2012.

[110] Y. Wang, M. T. Orchard, and A. R. Reibman, "Multiple description image coding for noisy channels by pairing transform coefficients," in *Proceedings of First Signal Processing Society Workshop on Multimedia Signal Processing*, pp. 419–424, Jun 1997.

[111] M. T. Orchard, Y. Wang, V. Vaishampayan, and A. R. Reibman, "Redundancy rate-distortion analysis of multiple description coding using pairwise correlating transforms," in *Proceedings of International Conference on Image Processing*, vol. 1, pp. 608–611 vol.1, Oct 1997.

[112] V. K. Goyal and J. Kovacevic, "Optimal multiple description transform coding of gaussian vectors," in *Data Compression Conference, 1998. DCC '98. Proceedings*, pp. 388–397, Mar 1998.

[113] A. R. Reibman, H. Jafarkhani, Y. Wang, M. T. Orchard, and R. Puri, "Multiple-description video coding using motion-compensated temporal prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, pp. 193–204, Mar 2002.

[114] L. Zhao, H. Bai, A. Wang, and Y. Zhao, "Multiple description convolutional neural networks for image compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2494–2508, 2019.

[115] L. Zhao, H. Bai, A. Wang, and Y. Zhao, "Deep multiple description coding by learning scalar quantization," in *2019 Data Compression Conference (DCC)*, pp. 615–615, 2019.

[116] U. Samarawickrama, J. Liang, and C. Tian, "M -channel multiple description coding with two-rate coding and staggered quantization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, pp. 933–944, July 2010.

[117] D. Wang, N. Canagarajah, and D. Bull, "Slice group based multiple description video coding using motion vector estimation," in *2004 International Conference on Image Processing, ICIP '04.*, vol. 5, pp. 3237–3240 Vol. 5, Oct 2004.

[118] C.-C. Su, J. Yao, and H. Chen, "H.264/avc-based multiple description coding scheme," in *IEEE International Conference on Image Processing, ICIP 2007.*, vol. 4, pp. IV – 265–IV – 268, Sept 2007.

[119] T. Tillo, M. Grangetto, and G. Olmo, "Redundant slice optimal allocation for h.264 multiple description coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, pp. 59–70, Jan 2008.

[120] S. Tian and K. P. Rajan, "Multiple description coding using transforms and data fusion," in *International Conference on Information Technology: Coding and Computing, ITCC 2005*, vol. 1, pp. 85–90 Vol. 1, April 2005.

[121] M.-T. Lu, J.-C. Wu, K.-J. Peng, P. Huang, J. Yao, and H. Chen, "Design and evaluation of a p2p iptv system for heterogeneous networks," *IEEE Transactions on Multimedia*, vol. 9, pp. 1568–1579, Dec 2007.

[122] C.-W. Hsiao and W.-J. Tsai, "Hybrid multiple description coding based on h.264," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, pp. 76–87, Jan 2010.

[123] Z. Xu, Z. Lin, and A. Makur, "Multiple description image coding with hybrid redundancy," in *IEEE Asia Pacific Conference on Circuits and Systems, APCCAS 2006*, pp. 382–385, Dec 2006.

[124] J. Y. Chen and W. J. Tsai, "Joint temporal and spatial multiple description coding for h.264 video," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pp. 1273–1276, Sept 2010.

[125] J. Chen, J. Liao, H. Zeng, and C. Cai, "Multiple description coding for multiview video with adaptive redundancy allocation," in *2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 181–185, 2017.

[126] A. Norkin, A. Aksay, C. Bilen, G. Bozdagi, A. Gotchev, and J. Astola, "Schemes for multiple description coding of stereoscopic video," pp. 11–13, 2006.

[127] H. Karim, C. Hewage, S. Worrall, and A. Kondoz, "Scalable multiple description video coding for stereoscopic 3D," *IEEE Transactions on Consumer Electronics*, vol. 54, pp. 745–752, May 2008.

[128] R. Choupani, S. Wong, and M. Tolun, "Spatial multiple description coding for scalable video streams," *International Journal of Digital Multimedia Broadcasting*, vol. 2014, 2014.

[129] L. P. Kondi, "Transactions letters. a rate-distortion optimal hybrid scalable/multipledescription video codec," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 921–927, July 2005.

[130] E. Akyol, A. M. Tekalp, and M. R. Civanlar, "Scalable multiple description video coding with flexible number of descriptions," in *IEEE International Conference on Image Processing, 2005. ICIP 2005.*, vol. 3, pp. III–712–15, Sept 2005.

[131] H. Mansour, P. Nasiopoulos, and V. Leung, "An efficient multiple description coding scheme for the scalable extension of h.264/avc (svc)," in *2006 IEEE International Symposium on Signal Processing and Information Technology*, pp. 519–523, Aug 2006.

[132] M. Yu, X. Ye, R. Wang, F. Xiao, and G. Jiang, "New multiple description layered coding method for video communication," in *Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies, 2005. PDCAT 2005.*, pp. 694–697, Dec 2005.

[133] X. Li, B. Yin, and D. Kong, "Multiple description image coding for scalable and robust transmission over ip," in *The Sixth World Congress on Intelligent Control and Automation, 2006. WCICA 2006*, vol. 2, pp. 9996–10000, 2006.

[134] N. Franchi, M. Fumagalli, R. Lancini, and S. Tubaro, "Multiple description video coding for scalable and robust transmission over ip," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 321–334, March 2005.

[135] L. Favalli and M. Folli, "A scalable multiple description scheme for 3D video coding based on the interlayer prediction structure," *International Journal of Digital Multimedia Broadcasting*, vol. 2010, 2010.

[136] J. G. Apostolopoulos and M. D. Trott, "Path diversity for enhanced media streaming," *IEEE Communications Magazine*, vol. 42, pp. 80–87, Aug 2004.

[137] G. Moore, "Cramming more components onto integrated circuits," *Proceedings of the IEEE*, vol. 86, pp. 82–85, Jan 1998.

[138] J. Ostermann, J. Bormans, P. List, D. Marpe, M. Narroschke, F. Pereira, T. Stockhammer, and T. Wedi, "Video coding with h.264/avc: tools, performance, and complexity," *IEEE Circuits and Systems Magazine*, vol. 4, pp. 7–28, First 2004.

[139] F. Bossen, B. Bross, K. Suhring, and D. Flynn, "Hevc complexity and implementation analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, pp. 1685–1696, Dec 2012.

[140] E. Rahimi and C. Joslin, "Reliable 3D video streaming considering region of interest," *EURASIP Journal on Image and Video Processing*, vol. 2018, 06 2018.

[141] E. Rahimi and C. Joslin, "3D video spatiotemporal multiple description coding considering region of interest," in *Proceedings of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020)*, pp. 474–481, Jan 2020.

[142] T. O. Kvlseth, "Coefficient of variation: the second-order alternative," *Journal of Applied Statistics*, vol. 44, no. 3, pp. 402–415, 2017.

[143] J. D. Curto and J. C. Pinto, "The coefficient of variation asymptotic distribution in the case of non-iid random variables," *Journal of Applied Statistics*, vol. 36, no. 1, pp. 21–32, 2009.

[144] M. Mirahsan, G. Senarath, H. Farmanbar, N. D. Dao, and H. Yanikomeroglu, "Admission control of wireless virtual networks in hethetn ets," *IEEE Transactions on Vehicular Technology*, vol. 67, pp. 4565–4576, May 2018.

[145] M. Mirahsan, Z. Wang, R. Schoenen, H. Yanikomeroglu, and M. St-Hilaire, "Unified and non-parameterized statistical modeling of temporal and spatial traffic heterogeneity in wireless cellular networks," in *2014 IEEE International Conference on Communications Workshops (ICC)*, pp. 55–60, June 2014.

[146] H.-H. Institut, "H.264/avc reference software," June 2015. `http://iphome.hhi.de/suehring/tml/`, Accessed: 26-Dec-2021.

[147] F. Bossen, D. Flynn, and K. S. andKarsten Shring, "Jctvcsoftware manua," *Joint Collaborative Team on Video Coding (JCTVC) of ITUT SG16 WP3 and ISO/IEC JTC1/SC29/WG11*.

[148] C. Fehna, K. Schuur, I. Feldmann, P. Kauff, and A. Smolic, "Distribution of ATTEST test sequences for EE4 in MPEG 3DAV," in *ISO/IEC JTC1/SC29/WG11 MPEG02/-M9219, Doc. no. M9219*, Dec 2002.

[149] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in *ACM SIGGRAPH and ACM Trans. on Graphics*, pp. 600– 608, Aug 2004.

[150] "3D stereoscopic photography." , `http://3dstereophoto.blogspot.com/p/software.html`, Accessed: 24-Dec-2021.

[151] D. R. I. M. Setiadi, "PSNR vs SSIM: imperceptibility quality assessment for image steganography," *Multimed. Tools Appl.*, 2020.

[152] E. Rahimi and C. Joslin, "3D video multiple description coding considering region of interest," in *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 4: VISAPP, (VISIGRAPP 2017)*, pp. 208–215, 2017.

[153] K. Mansfield and J. Antonakos, *Computer networking from LANs to WANs*. Boston [Mass.]: Course Technology Cengage Learning, 2010.

[154] R. Pauliks, I. Slaidins, K. Tretjaks, and A. Krauze, "Assessment of ip packet loss influence on perceptual quality of streaming video," in *2015 Asia Pacific Conference on Multimedia and Broadcasting*, pp. 1–6, 2015.