DEVELOPMENT OF LOW COMPLEXITY ENCODER AND SUMMARIZATION TECHNIQUES FOR WIRELESS CAPSULE ENDOSCOPY VIDEO

Thesis

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

SUSHMA B

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA,

SURATHKAL, MANGALORE - 575025

APRIL 2022

DECLARATION

by the Ph.D. Research Scholar

I hereby declare that the Research Thesis entitled **DEVELOPMENT OF LOW COMPLEXITY ENCODER AND SUMMARIZATION TECHNIQUES FOR WIRELESS CAPSULE ENDOSCOPY VIDEO** which is being submitted to the **National Institute of Technology Karnataka, Surathkal** in partial fulfilment of the requirements for the award of the Degree of **Doctor of Philosophy** in **Electronics and Communication Engineering** is a *bonafide report of the research work carried out by me.* The material contained in this Research Thesis has not been submitted to any University or Institution for the award of any degree.

Sushma B
Reg. Number: 177124EC501
Department of Electronics and Communication Engineering
NITK Surathkal

Place: NITK, Surathkal.

Date:

CERTIFICATE

This is to certify that the Research Thesis entitled **DEVELOPMENT OF LOW COMPLEXITY ENCODER AND SUMMARIZATION TECHNIQUES FOR WIRELESS CAPSULE ENDOSCOPY VIDEO** submitted by **SUSHMA B** (Register Number: 177124EC501) as the record of the research work carried out by her, is accepted as the *Research Thesis submission* in partial fulfillment of the requirements for the award of degree of **Doctor of Philosophy**.

Dr. Aparna P
Research Guide
Dept. of E & C Engg.
NITK Surathkal - 575025

Chairman - DRPC
Department of Electronics and
Communication Engineering
(Signature with Date and Seal)

Acknowledgements

First and foremost, I would like to thank God, the almighty, who has granted countless blessing and for giving health and strength to my family and me while carrying out the research work.

I would like to express my sincere gratitude to my research guide Dr. Aparna P, Assistant Professor, Department of Electronics and Communication Engineering (ECE), NITK, Surathkal, for her guidance, endless support, encouragement and kindness throughout my research work. Mainly, I would like to thank her for her enormous patience in correcting my research articles and reports. For all these, I sincerely thank her from bottom of my heart and will be indebted to her throughout my life time.

I am grateful to Prof. John D' Souza for giving me an opportunity to pursue Ph.D at NITK. I would also express my gratitude to Prof. U. Shripathi Acharya, Head of the ECE department, during my admission for the Ph.D. program, Prof. T Laxminidhi and Prof. Ashvini Chaturvedi, Heads of the Department of ECE during my research work for their support, help, and encouragement.

I express heartfelt thanks to my Research Progress Assessment Committee (RPAC) members Dr. Raghavendra B.S., Dept. of ECE. and Dr. Geetha V., Dept. of Information Technology, for their valuable suggestions and constant encouragement to improve my research work. I sincerely thank all teaching, technical and administrative staff of the Electronics and Communication Engineering for their help during my research work.

I express my gratitude to the Ministry of Human Resource Development (MHRD) Government of India and TEQIP-II for providing financial assistance during my research. I would like to thank Dr. Raj Vigna Venugopal, Chief Gastroenterologist and staff of gastroenterolgy department, Manipal Hospitals, Bangalore for providing capsule endosocpy videos and valuable inputs to carry out this research. I would also

like to thank the research scholars of NITK who were always ready to help and open for the discussion. I would like to express my gratitude to everyone who assisted me in completing this thesis in one way or the other.

I would like to thank my parents for their support and encouragement. Surely, this research work would not have been possible without my husband and son's support. I greatly value their patience, support and understanding during my pursuit of this research work.

Place: Surathkal Sushma B

Date:

Dedicated to

My Family

Abstract

Wireless capsule endoscopy (WCE) is the state-of-the-art medical procedure for scanning the entire digestive tract to diagnose gastrointestinal (GI) diseases. Its non-invasiveness and ease of usage make it a better option than conventional endoscopy. However, it is inferior to conventional endoscopy due to low image quality imposed by capsule's complexity and power consumption. In one complete scan of GI tract, a capsule captures between 90000 and 180000 frames during its peristalsis movement. Diagnosing such a large number of images is a time-consuming and tedious procedure that needs a gastroenterologist's undivided attention. The main aim of the research work is two folds. One involves the development of a low complexity video encoder that can reduce the computations in the capsule. The other part involves a WCE video summarization framework to provide an efficient diagnosis.

Developing a low-complexity video encoding architecture that can achieve high compression performance at a low bit rate while maintaining acceptable reconstruction quality is a challenging task in WCE. A distributed video coding (DVC) architecture is proposed to achieve this, which transfers encoder complexity to the decoder side. It employs a keyframe encoder that takes advantage of GI image textural properties to reduce the complexity. Furthermore, the low-frequency bands of the Wyner-Ziv (WZ) frames are used as auxiliary information at the decoder to generate high-quality side information that enables the encoding of high frequency bands with a low bit rate. The proposed DVC framework is further enhanced to reduce complexity by eliminating WZ-chroma component encoding. Exploiting the similarity in colour and texture properties between consecutive frames in WCE video, a deep convolutional neural network model is integrated into the decoder side to predict the chroma component. The proposed methods achieve improvements in coding gain with low complexity encoder when compared with benchmark compression schemes.

A physician must dedicate lot of time in reviewing the large number of frames, and there is a considerable risk of missing frames that are associated with lesion symptoms. Review time can be minimized by extracting the summary of WCE video by eliminating the redundant frames. To achieve this, a summarization framework consisting a shot boundary detection and keyframe extraction methods is presented.

The proposed framework achieves better summarization performance measured using F-score and compression ratio compared to state of the art WCE summarization methods.

Keywords: Wireless capsule endoscopy, Video compression, Distributed video coding, Convolutional neural networks, Chroma prediction, Video summarization.

Contents

List of Figures

List of Tables

ABBREVIATIONS

BT-KFE	Block Textured conditioned Keyframe Encoder
CNN	Convolutional Neural Network
CR	Compression Ratio
DCT	Discrete Cosine Transform
DI	Diagnostic Yield
DPCM	Differential Pulse Code Modulation
DTT	Discrete Tchebichef Transform
DVC	Distributed Video Coding
HOG	Histogram of Gradients
GLCM	Gray Level Co-occurance Matrix
GOP	Group of Pictures
HEVC	High Efficiency Video Coding
JPEG	Joint Photographic Experts Group
LBP	Local Binary Pattern
LDPCA	Low Density Parity Check and Accumulate
MSE	Mean Square Error
NMF	Non Matrix Factorization
OBME	Overlapped Block based Motion Estimation
PSNR	Peak Signal to Noise Ratio
RD	Rate Distortion
RF	Radio Frequency
SI	Side Information
SNN	Siamese Neural Network
SSIM	Structural Similarity Index
SVM	Support Vector Machine
WZ	Wyner-Ziv

Chapter 1

Introduction

Endoscopy has become a standard and the most preferred method by physicians for detecting gastrointestinal (GI) disorders such as gastric cancer, polyps, intestinal bleeding, Crohn's disease and Celiac disease. It enables the direct visualization of the human GI tract and can even detect early cancers. Wired endoscopy is a commonly used procedure to diagnose the upper GI tract. However, the patients hesitate to undergo this procedure because of the pain and discomfort induced by inserting a long, flexible wire into the digestive tract. Moreover, the wired endoscopy cannot scan the small intestine due to its intricate and curvy nature.

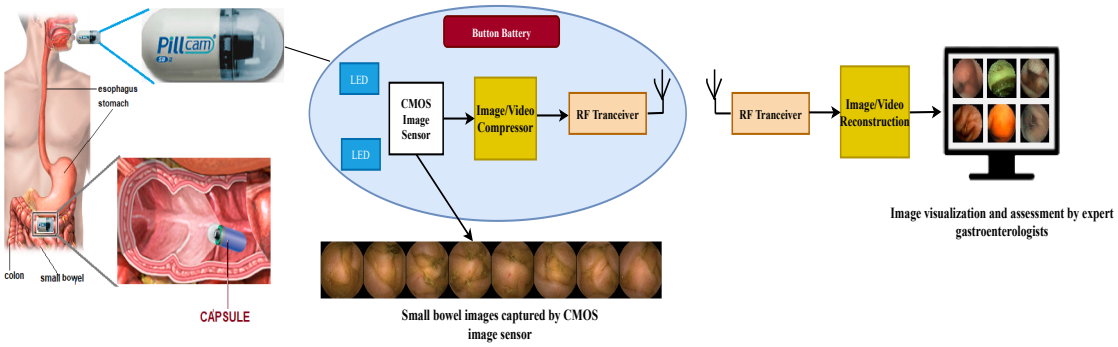


Figure 1.1: A typical WCE based GI tract screening process

Wireless capsule endoscopy (WCE) is used in recent days in order to overcome the drawbacks of wired endoscopy. In WCE, the patient swallows an electronic capsule that moves through the GI tract by peristalsis action. The capsule scans the entire GI tract including small intestine capturing its images without any discomfort and pain (Iddan *et al.* (2000), Wang *et al.* (2013)). The images captured during the

peristalsis movement of the capsule are transferred wirelessly through radio frequency (RF) transmission to an external recording unit. Later, the image data is transmitted to a computer and reconstructed before being analysed for GI abnormalities. A typical WCE based GI tract screening process is shown in Figure 1.1. Prior to the introduction of WCE, it was not possible to diagnose the small intestine without surgery.

Table 1.1: FDA approved clinically used capsule endoscopes.

Capsule	Inventor	Weight (g)	Dimension (mm)	Frame rate	Resolution	Angle of View	Capsule Life
Pill-Cam SB3	Given Imaging	3.4	11.0 x 26.0	2-6 fps	256 x 256	160°	10-12h
Endo-Capsule	Olympus	3.8	11.0 x 26.0	2 fps	256 x 256	160°	8-10h
Miro-Cam	Intromedic	3.4	11.0 x 24.0	2 fps	320 x 320	170°	8-11h
OMOM	Jinshan	6.0	28.0 x 13.0	2 fps	320 x 320	140°	7-9h
Navi-Cam	Ankon Tech.	5.0	28.0 x 12.0	2 fps	480 x 480	140°	8-9h

Many commercial swallowable capsule endoscopes have been developed to assist in the non-invasive scanning of the complete GI tract. The most common clinically used capsule endoscopes approved by the Food and Drug Administration (FDA) available are listed in Table 1.1, with their specifications. To make the transition through the GI tract easier, the capsules are made with a smaller diameter and length. Most of the commercially available WCE systems operate at modest frame resolutions of 256x256 pixels and a frame rate of 2-6 fps on the battery life of 7-8 hours (Gurudu *et al.* (2008), Ciuti *et al.* (2011)). Even with many benefits of WCE, this technology is still considered immature due to low diagnostic yield (DI) and the prolonged labour-intensive diagnosis procedure. DI can be improved by increasing frame resolution, frame rate, and working duration to achieve quality diagnoses. High frame resolution is essential for diagnosing early-stage lesions. When the images are zoomed in for an in-depth diagnosis, the low-resolution images fail to describe the finer details of the lesions. A higher frame rate enables the identification of more lesions and reduces the chance of missing frames containing significant information. But, improving the resolution and frame-rate increases the image data for processing and transmission. Around 70% to 90% of the power is consumed by RF transmission (Chen *et al.*

(2009)), which reduces the capsule battery life. However, the capsule battery must have enough power to run for more than 16 hours in order to successfully complete the entire GI tract scan.

One complete examination of the WCE procedure captures thousands of images of different parts of the GI tract. Transmitting and processing these images at a high resolution and frame rate consumes a lot of power. In addition, manual assessment of a large number of frames is a tedious task which requires a lot of attention from the doctor and is prone to more diagnostic errors (Hernandez-Lara and Rajan (2021), Zheng *et al.* (2012)). Diagnostic accuracy is crucial in detecting the abnormalities and entirely depends on the WCE video reviewer's expertise. The size of the video data needs to be reduced without compromising its quality before transmission to achieve better DI with low power consumption, which can be achieved by video compression (Ou *et al.* (2015)). When it comes to video compression techniques, it is apparent that wireless capsule endoscopes experience major limitations in terms of processing capability and power consumption. Therefore, low complexity compression algorithms must be used to avoid the capsule being bulky and the image compressor itself consuming additional power.

Furthermore, because the recorded video is used for medical diagnostics, providing high quality decoded video with a high compression ratio is essential. In WCE, many frames with high similarity exist due to slow movement and sometimes no-movement of the capsule in some areas of GI tract (Barducci *et al.* (2020)). Similar frames introduce much redundancy in the WCE video. Hence, developing an algorithm to generate the WCE video summary to remove the redundancy is important. Removing redundant frames saves the time required for the laborious task of assessing the video to find the abnormalities, making WCE a more efficient diagnosis procedure.

1.1 Motivation

Many standard video compression designs such as MPEG standards, H.264/AVC and H.265/HEVC exist that focus on obtaining optimized compression performance. The standard video coding architectures compress the video multiple times by removing the redundancy at the encoder using inter frame motion prediction. These systems

employ joint encoding and decoding, where the encoder's computational complexity is very high and that of the decoder is low. The high computational complexity of the encoder is due to the motion estimation performed to remove the temporal correlation existing between the consecutive frames in a video sequence to achieve high compression performance. Hence, the standard video compression architectures are unsuitable for attaining efficient video compression with the low power consumption requirements set by the capsule endoscope. Developing an efficient video compression method with a low complexity encoder in the context of capsule endoscopy with limited processing capability and battery life is a challenging problem. The principle of DVC enables the development of a low complexity video encoding architecture to reduce processing and transmission power consumption.

In recent years, a lot of effort has been made into developing computer-aided diagnosis methods to assist doctors in reducing the time and manpower required for WCE video examination. The capsule captures the images at the rate of 3 to 6 frames per second for over 8 hours and acquires around 90000-180000 frames. A physician has to invest a lot of time or appoint an assistant to inspect these huge number of frames and summarize the endoscopy video by eliminating redundant frames. The major disadvantage associated with manual summarizing is the chance of eliminating some of the frames with lesion symptoms while inspecting thousands of images. Some of the methods are proposed for the detection of lesions which includes tumours, ulcers, polyps and Crohn's disease (Jia *et al.* (2019), Klang *et al.* (2020)). A few approaches are proposed for the detection of lymphangiectasias, Celiac disease and hookworms (Li *et al.* (2019), He *et al.* (2018)). All these methods deal with detecting only one or two types of abnormalities. The majority of the frames with other abnormalities still need to be manually assessed by the gastroenterologist. To overcome all the above drawbacks, developing an algorithm to generate video summary without missing frames with sensitive information is very crucial. Therefore, video summarization is considered as the best approach to reduce the review time which provides a comprehensive view of the entire WCE video. WCE video summarization tool allows the physician to get a quick glimpse of the overall content in the video and the presence of possible abnormalities using a summarized video consisting of only keyframes. Any frames with sensitive information are found, the physician can always refer to the adjacent

frames in the original video.

Motivated by low complexity encoder requirements for WCE video application and the importance of an efficient WCE video summarization system, the objectives of the research work are formulated as follows:

- **Design of low complexity video compression algorithm suitable for WCE and a decoder for high quality reconstruction of the compressed video.**

- **WCE video summarization framework to eliminate the redundancy without losing the significant frames.**

1.2 Contributions of the Research

The main focus of this research is to develop strategies to reduce the video encoder complexity of the capsule while maintaining the compression performance and quality of reconstruction. The work also investigates the techniques for the WCE video summarization. The key contributions of the research are as follows:

- Proposed a DVC-based low complexity video encoder to solve encoder complexity constraints in the WCE video application by exploiting the degree of freedom available for complexity at the decoder.

- Developed a keyframe encoding method, which exploits the textural characteristics of WCE images to reduce the computations required for processing by differentiating the keyframe blocks into smooth and non-smooth transformed blocks. The quality of side-information (SI) production at the decoder determines the compression performance of DVC. By dividing the WZ frequency coefficients into intra and WZ bands, a method for generating high-quality SI is proposed.

- Presents a method where only the luma component of the WZ frame is processed and encoded. The chroma component of the WZ frame is predicted by a CNN based deep chroma prediction model. The model is trained to predict chroma by matching luma and texture information of the keyframe and WZ frame at the

decoder. The proposed method reduces the computational complexity required for encoding the WZ chroma component.

- An unsupervised WCE video summarization framework is proposed consisting of deep feature extraction, video shot detection and keyframe selection. Deep features extracted by convolutional autoencoder are used to segment the video into shots based on the similarity between the frames. A technique to construct a motion profile to extract keyframes from each shot is proposed to generate the final WCE video summary.

1.3 Simulation Environment and Datasets

1.3.1 Simulation Environment

The presented results of the proposed and benchmark methods are computed using an Intel core i5-7200 2.5GHz CPU, 8GB RAM and NVIDIA GeForce 940MX GPU. Implementation of the proposed compression system is done in MATLAB and the performance is evaluated against MJPEG, TDWZ-DVC and H.264/AVC-Intra codecs. Deep neural network implementation and training is done using Keras, a deep learning API using Tensorflow as backend on an NVIDIA Tesla-T4 GPU.

1.3.2 Datasets

The WCE video compression system is evaluated on four test video sequences captured by Mirocam-Intromedic capsule with a frame resolution of 320 x 320 at different organs of GI tract. These videos are collected from Department of Gastroenterology, Manipal Hospitals, Bangalore, India. The details of the test video sequences is given

Table 1.2: Test video sequences used to evaluate the WCE video compression system

Test Video Sequence	Captured GI organ	Motion Type	Video frame length
Video-1	Small intestine	No and very less motion	400
Video-2	Stomach	Slow to Moderate motion	350
Video-3	Esophagus	Fast motion	280
Video-4	Colon	Moderate to fast motion	300

in Table 1.2. For evaluating the performance of the WCE video summarization technique, keyframes of around 4 video sequences captured at different locations of GI and 3 video sequences of KID-dataset are identified with the help of gastroenterologist. WCE video summarization dataset details are given in Table 1.3. All the sequences in the dataset are captured by Intromedic-Mirocam capsule with a frame resolution of 320 x 320.

Table 1.3: Test sequences used to evaluate the WCE video summarization performance

Test Video Sequence	Captured GI organ	Video frame length	Number of Keyframes	Source
KID-1	All GI organs	65000	12520	**KID Dataset-1**
KID-2	All GI organs	62000	11700	MDSS research group
KID-3	All GI organs	62500	11904	(KID Dataset (2017))
Video-1	Small intestine	5922	600	**Dataset-2**
Video-2	Colon	3000	590	Manipal Hospitals,
Video-3	Esophagus	390	150	Bangalore
Video-4	Stomach	450	135	

1.4 Evaluation Parameters

1.4.1 Video Compression

The evaluation of the compression system is done by comparing compression ratio (CR) calculated using (1.1) with the quality metrics PSNR and structural similarity index (SSIM). PSNR in dB and SSIM determine the amount of quality lost in encoding process and computed using (1.2) and (1.4) respectively.

$$CR = \left(1 - \frac{\text{Total bits after compression}}{\text{Total bits before compression}}\right)\text{x100} \tag{1.1}$$

$$PSNR = 10\log_{10}\frac{255^2}{MSE} \tag{1.2}$$

where

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}|x(i,j) - x_r(i,j)|^2 \tag{1.3}$$

where x and x_r are original and reconstructed pixel intensities of an MXN image.

$$SSIM(x, x_r) = \frac{(2\mu_x\mu_{x_r} + C_1) + (2\sigma_{xx_r} + C_2)}{\left(\mu_x^2 + \mu_{x_r}^2 + C_1\right)\left(\sigma_x^2 + \sigma_{x_r}^2 + C_2\right)}, \tag{1.4}$$

where $C_1 = 6.5, C_2 = 58.52$ are constants, μ_x, μ_{x_r} are mean intensities and σ_x, σ_{x_r} are standard deviations of x and x_r respectively. σ_{xx_r} is computed using the following equation.

$$\sigma_{xx_r} = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \mu_x\right)\left(x_{r_i} - \mu_{x_r}\right) \tag{1.5}$$

Other important parameters considered to evaluate a video coding system's performance over reference video coding systems are Bjontegaard-Delta (BD) metrics (Bjontegaard (2001)). BD-metrics are widely used to compare a video coding system's performance with the reference codec over a range of quality points or bitrates. BD metric is often computed as change in bitrate or a change in quality measured using PSNR and SSIM based on rate-distortion (RD) curves from the tested data points. The BD-rate represents the average bitrate savings for the same video quality and is calculated between RD curves of the tested video codec A and a reference codec B. The bitrate saving difference between the two RD curves belong to codecs A and B at a given PSNR is computed by (1.6). BD-PSNR and BD-SSIM between two RD curves A and B at a given bitrate is computed by (1.7) and (1.8).

$$BD - Rate = \frac{Rate_A(PSNR) - Rate_B(PSNR)}{Rate_B(PSNR)} \tag{1.6}$$

$$BD - PSNR = \frac{PSNR_A(bitrate) - PSNR_B(bitrate)}{PSNR_B(bitrate)} \tag{1.7}$$

$$BD - SSIM = \frac{SSIM_A(bitrate) - SSIM_B(bitrate)}{SSIM_B(bitrate)} \tag{1.8}$$

1.4.2 Video Summarization

Video summarization performance is evaluated by using F-score computed using (1.9) which is a function of precision (p) and recall (r) computed using (1.10) and (1.11) respectively.

$$F\text{-score} = \frac{2rp}{r+p} \tag{1.9}$$

$$p = \frac{TP}{TP + FP} \tag{1.10}$$

$$r = \frac{TP}{TP + FN} \tag{1.11}$$

- TP (True-positive) is the number of correct matches between keyframes extracted from proposed method and ground-truth summary.
- FN (False-negative) is the number of frames which are in the result but not present in ground-truth summary.
- FP (False-positive) is the number of frames in the ground-truth but not in result.

Compression ratio (CR) is calculated using (1.12), where N_k is number of WCE keyframes extracted using proposed method and N_t is total number of frames in a video sequence.

$$CR = 1 - \frac{N_k}{N_t} \tag{1.12}$$

1.5 Outline

The remainder of the thesis is structured as follows:

- **Chapter 2** describes the peristalsis behaviour of the GI tract to understand the capsule's speed and time spent in different parts of the GI tract, as well as the colour and texture features of the images captured in the GI tract. It provides a comprehensive review of prior literature on WCE image compression and WCE video summarization techniques. The first part of the chapter mainly focuses on the existing WCE image compression techniques and the need for the compressor to exploit temporal correlation. It also discusses the limitations of standard video compression architectures in designing low complexity encoders. A detailed description of H.264-Intra, TDWZ-DVC and MJPEG encoding methods used as reference encoding schemes to evaluate the performance of the WCE compression systems is presented. The chapter describes the video summarization technique, the necessity to summarize the WCE video and the review of the prior work.

- **Chapter 3** presents a DVC architecture suitable for WCE video compression, which encodes the WZ frames by classifying frequency bands into intra and WZ bands. A technique to reduce the computations of the JPEG based keyframe encoder by exploiting the textural characteristics is presented. It also explains the modification done at the transform and quantization blocks to reduce the complexity. Also, a new way of WZ encoding of subsampled chroma components is presented in the chapter. The proposed method is evaluated by comparing RD performance and encoding complexity with the MJPEG, TDWZ-DVC and H.264-Intra methods.

- **Chapter 4** presents the improvisation of the DVC architecture proposed in Chapter 3. The DVC architecture with deep chroma prediction model incorporated at the decoder to reduce the complexity of WZ frame chroma components encoding is presented. The deep chroma prediction model's architectural details, loss function and the training details are provided. The chapter presents the significance of colour space in training the model. The presented DVC method eliminates the complexity of the encoder required for chroma encoding and achieves better RD performance compared to the method presented in Chapter 3, MJPEG and TDWZ-DVC at the reduced complexity.

- **Chapter 5** introduces a framework to obtain a summary of WCE video using deep feature matching and motion analysis. A method for deep features extraction using a convolutional autoencoder and a method for segmenting the WCE video into shots by using the deep features is presented. The motion profile creation for a video shot and the method to extract the most representative frames from each shot using the motion profile is presented. The performance of the presented method is evaluated by using F1-Score and compression ratio.

- **Chapter 6** provides an overall conclusions of the research and future directions.

Chapter 2

Background

The structure and peristalsis behaviour of GI tract and analysis of the GI tract image characteristics is necessary to develop an efficient WCE imaging system. This chapter presents the capsule's speed and time it spends in each organ, the colour and texture characteristics of the images captured in various organs of the GI tract. Further, this chapter provides a comprehensive review of the existing WCE image and video compression, as well as summarization techniques.

2.1 Structure and Peristalsis Behaviour of GI tract

The capsules are designed to travel through the GI tract organs of different structure through peristalsis actions. Therefore, it is important to understand the structure and peristalsis actions of different organs of the GI tract to create an efficient WCE imaging algorithm. A general description of the structure and peristalsis behaviour of the GI tract is given in this section. The human GI tract consists esophagus, stomach, small bowel and large intestine which are tubular structured organs connected in series. The total approximate length of the GI tract is around $800cm$ - $900cm$. Esophagus is a long tube-like structure, which tries to propel the capsule towards the stomach with peristaltic actions. The stomach is a J-shaped organ that tumbles to mix the food and liquid secreted. Due to this action of the stomach, images captured in the stomach exhibits an irregular motion. The small intestine is a crucial organ of the GI tract which can be directly visualized only using the WCE procedure. It is a narrow, curvy and long tubular like structure where the capsule exhibits very slow motion. The large intestine, also called the colon, is the last part of the GI tract which is

broader and shorter than the small intestine.

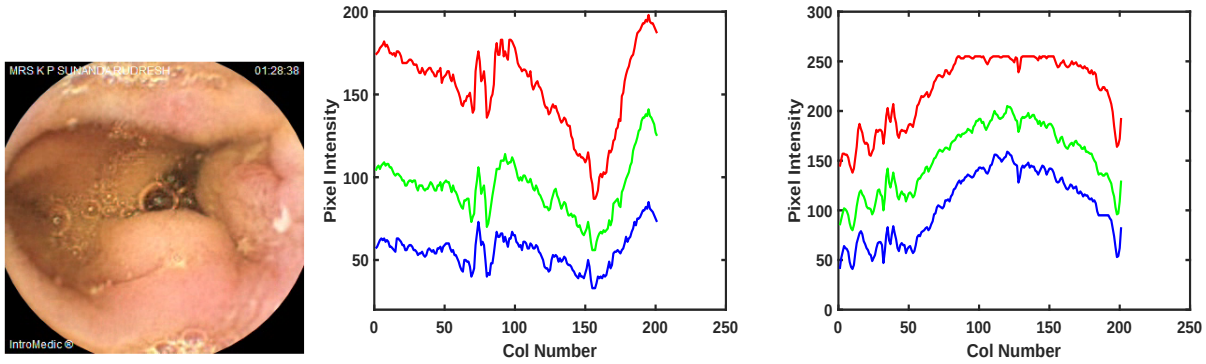
Table 2.1: GI tract structure and capsule movement details (Liu *et al.* (2015))

GI organ	Organ Length (cm)	Capsule speed (mm/s)	Motion in images	Transit time
Esophagus	18 - 25	6 - 10	Fast	18 - 50 sec
Stomach	—	3 - 8	Moderate to Slow	100 - 120 min
Small intestine	500 - 650	0.2 - 1	Slow or no motion	241 - 402 min
Large intestine	155 - 170	1 - 10	Moderate to Slow	60 - 100 min

Table 2.1 provides the details of the length of the organs, the capsule's speed in various GI organs, the type of motion exhibited in the images captured, and the amount of time spent in each organ. These parameters are vital in understanding the capsule motion to design efficient WCE imaging algorithms. The capsule spends a significant amount of the time in the small intestine because of the slow movement and the images captured exhibit very low or no motion.

2.2 Analysis of WCE Image Characteristics

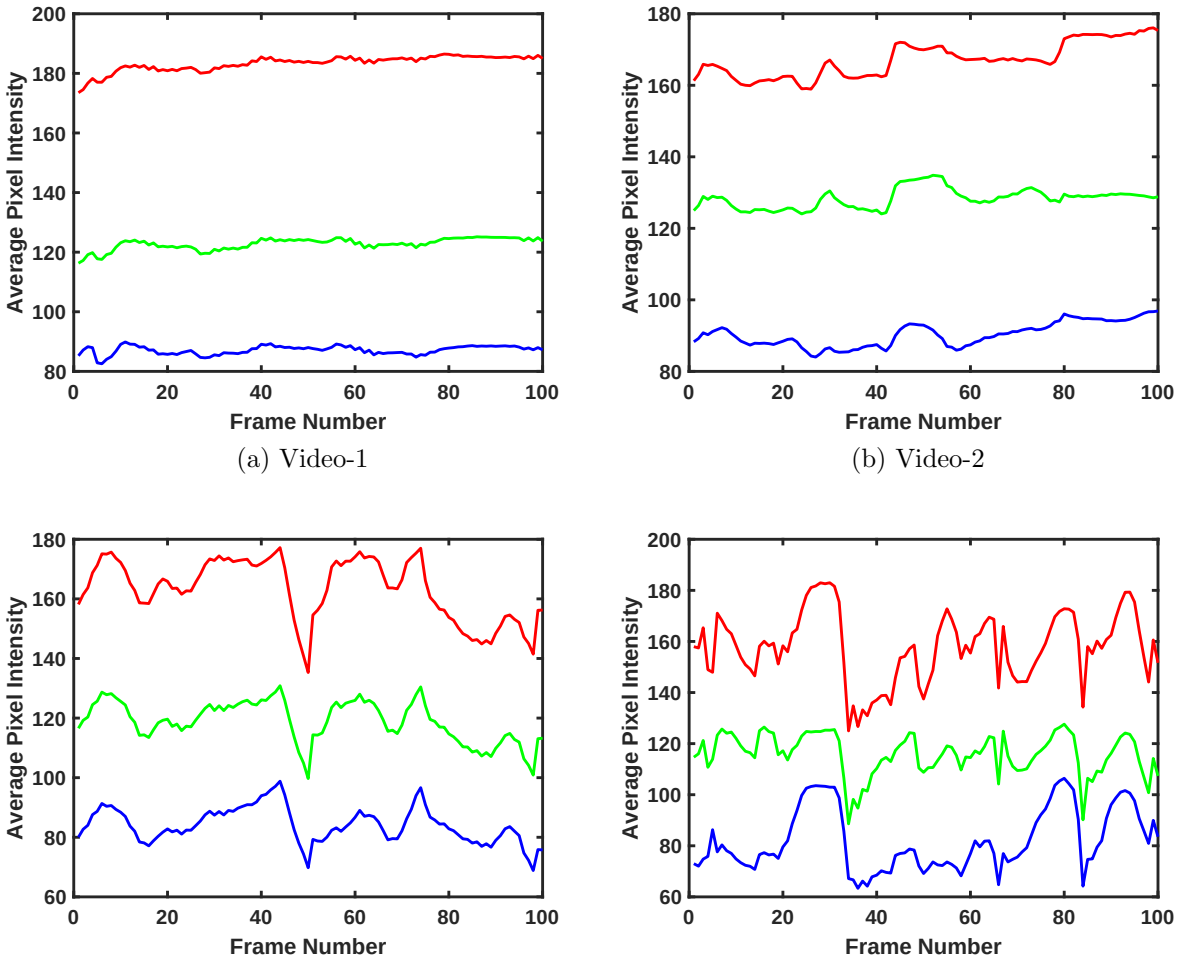
2.2.1 Colour Space Conversion



(a) Img1 (b) img1: Row 100 (c) img1: Row 200

Figure 2.1: WCE Images and their RGB profile along rows 100 and 200

The image sensor used in the capsule captures the images in RGB colour space while moving through the GI tract. Human GI system generally looks red except in abnormal regions. Green and blue components are less significant compared to red

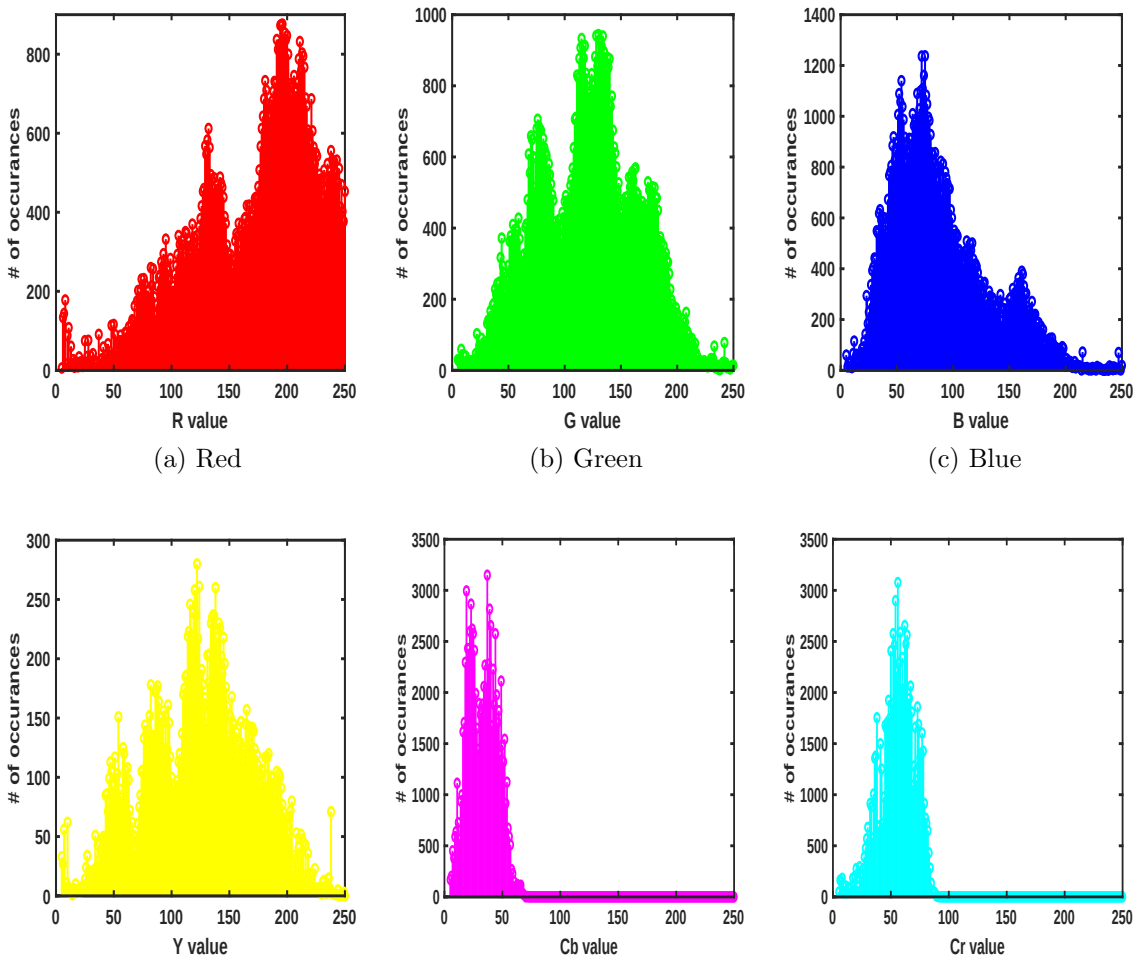


(a) Video-1 (b) Video-2

(c) Video-3 (d) Video-4

Figure 2.2: Mean Intensity of RGB components for frames of different video sequences

component. On an average, in most of the GI endoscopic images, the red component pixel value is the highest and blue component is the lowest as depicted in the Figure 2.1. The mean intensity distribution of RGB colour components of different frames of test video sequences is given in Figure 2.2. Generally, in most of the colour image and video compression techniques, RGB colour space is transformed into another colour space using reversible colour transformation to de-correlate the colour components. A reversible colour space transformation is introduced to transform WCE images captured in RGB colour space to YCbCr colour space at the encoder as given in (2.1). In YCbCr, Y represents the luminance component and Cb, Cr components stores the chrominance components. At the decoder, the YCbCr components are converted to RGB components using inverse colour space transformation given



(a) Red (b) Green (c) Blue

(d) Luma-Y (e) Chroma-Cb (f) Chroma-Cr

Figure 2.3: Histogram of R, G, B, Y, Cb and Cr components of an endoscopic image

in (2.2).

$$Y = \frac{R}{4} + \frac{G}{2} + \frac{B}{4}, \quad Cb = B - G, \quad Cr = R - G \tag{2.1}$$

$$R = Y - \frac{Cb}{4} + \frac{3Cr}{4}, \quad G = Y - \frac{Cb}{4} - \frac{Cr}{4}, \quad B = Y - \frac{Cr}{4} + \frac{3Cb}{4} \tag{2.2}$$

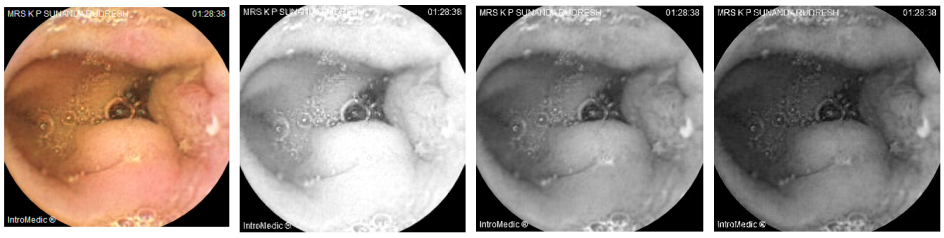
As we can realize from histogram plots shown in Figure 2.3, in RGB color space all the three components have high variation in pixel values. However, in YCbCr color space the variation is quite low in Cb and Cr components. The role of color space conversion in image compression algorithms is to reduce the high frequency contents in the image. The relative frequency content of an image in a spatial domain can be estimated using standard deviation. Table 2.2 shows the average standard deviation and entropy (in bits/pixel) values for different video sequences of 100 frames.

The reduction in standard deviation can be observed in case of YCbCr colour space compared to RGB colour space.

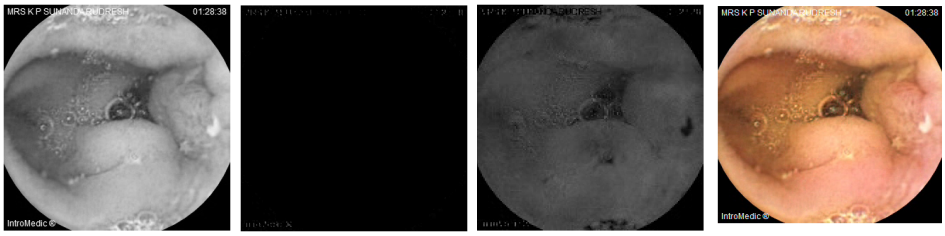
Table 2.2: Statistical measurements of colour space components

Video Sequence	Statistical measurements	Colour Channels					
		R	G	B	Y	Cb	Cr
Video 1	Average standard deviation	73.81	60.06	48.78	59.56	25.07	20.58
	Average entropy (bits/pixel)	7.08	6.48	6.99	7.15	5.899	5.67
Video 2	Average standard deviation	73.84	52.09	40.37	53.57	27.32	19.98
	Average entropy (bits/pixel)	6.96	6.76	6.63	6.89	5.88	5.39
Video 3	Average standard deviation	71.94	48.62	38.06	50.73	28.35	18.11
	Average entropy (bits/pixel)	6.98	6.67	6.48	6.82	5.91	5.23
Video 4	Average standard deviation	72.56	59.21	47.67	58.21	25.32	20.19
	Average entropy (bits/pixel)	7.02	6.08	6.72	6.99	5.79	5.67

2.2.2 Subsampling of Chroma Components



(a) Original (b) Red (c) Blue (d) Green



(e) Luma-Y (f) Chroma-Cb (g) Chroma-Cr (h) Reconstructed

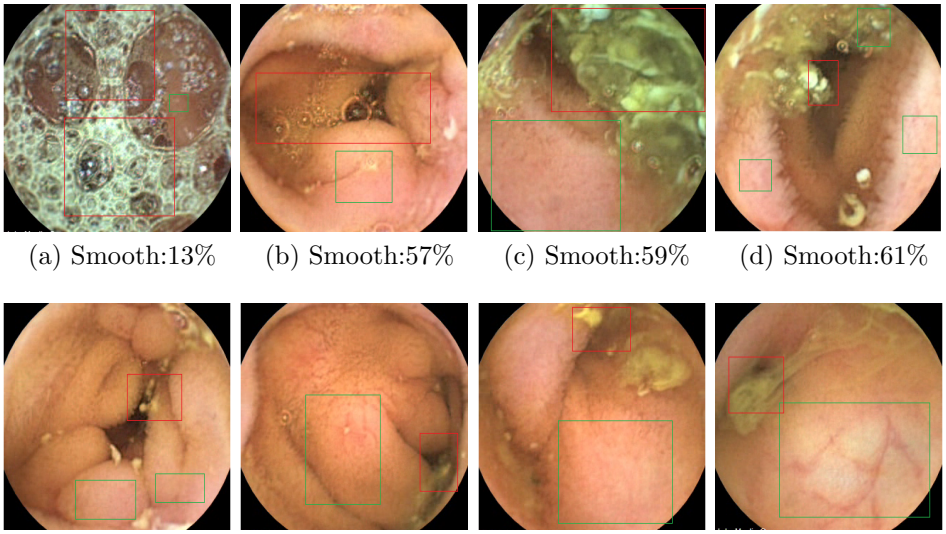
Figure 2.4: Original and reconstructed image after chroma subsampling along with R, G, B and Y, Cb, Cr components of a WCE image

In WCE video frames, chroma pixel intensities are almost same with in a small region and can be down-sampled without loss of significant information in order to

reduce the size of the image to be encoded. Original image, RGB, YCbCr components and reconstructed subsampled chroma components at 4:2:0 format are shown in Figure 2.4. Sub-sampling is a simple and efficient method which has been successfully used in endoscopic image compression algorithms. The 4:2:0 sub-sampling format is used in compression of endoscopic images where for every 4 pixels of Y-component, 1 pixel of Cb and Cr are selected. The average reconstruction quality of the upsampled images is around 50dB in PSNR and 0.9998 in SSIM.

2.2.3 Smooth Blocks and Textured Blocks

Block classification method to classify blocks into smooth and non-smooth blocks can be incorporated to reduce the complexity of quantization and entropy coding. Most of the region in WCE images is smooth in nature with uniform pixel intensities. This kind of blocks have significant energy only in low frequency components. WCE images



(a) Smooth:13% (b) Smooth:57% (c) Smooth:59% (d) Smooth:61%

(e) Smooth:71% (f) Smooth:72% (g) Smooth:79% (h) Smooth:85%

Figure 2.5: Percentage of smooth blocks of size 8x8 in various WCE images

with percentage of smooth blocks is shown in Figure 2.5. Smooth and non-smooth regions are represented using green and red rectangles respectively. The average percentage of smooth blocks considered in the test video sequences are 70%.

2.3 Literature Review on Compression in WCE

The WCE procedure is performed mainly to diagnose the GI tract in order to identify the abnormalities. Images provided for diagnosing should be of high quality. However, most of the commercial wireless capsules available work at a frame rate of 2-6 frames per second (fps) with a moderate frame resolution of 256x256 pixels (Alam *et al.* (2017)). In most of the cases, these quality standards are insufficient for a thorough and accurate diagnosis. The power consumption increases as the image resolution and frame rates are increased. To minimize the power consumption, image compressor should be used to reduce the size of the image transmitted without losing important information.

2.3.1 Challenges in WCE Video Compression

Designing a low complexity image compressor with high compression and reconstruction quality that consumes significantly less power is a challenging problem. According to many literature studies, compressors suitable for WCE should abide by the following principles.

- Extensive image processing cannot be performed within the data compressor designed for WCE due to low power supply. So, each pixel can use very few simple operations.
- Compressor should be memory efficient as memory access consumes more power that increases the complexity of the compressor. Storing large amounts of data during processing requires more memory stores and loads. Therefore, the compression algorithm designed should be able to process the data with minimal intermediate storing.
- To detect some lesions, zooming of the image is required which will introduce blurriness in the low resolution image and there are greater chances of missing out on the significant lesions. Higher resolution is required for an accurate diagnosis. Therefore studies suggest to have minimum 512 x 512 image resolution.
- An observational study suggests more frames per second detect more features and lesions, reducing the rate of significant information loss. High reconstructed

image quality measured in peak signal to noise ratio (PSNR) is required for high quality images. According to a study (Istepanian *et al.* (2008)), acceptable PSNR required for medical images is minimum 35dB.

2.3.2 WCE Image Compression Methods

A basic WCE image compression system includes colour space conversion, prediction or transformation, quantization, and entropy encoding. A colour space conversion from RGB to YCbCr colour space is widely used in compressing endoscopic images as discussed in Section 2.2.1. Images can be encoded in time domain or transform domain. Time domain encoding uses predictive algorithms such as differential pulse code modulation (DPCM) and JPEG-LS which are lossless or near-lossless. In transform domain coding, the image is divided into block of pixels where each block is decomposed into set of frequency coefficients. Most of the low frequency coefficients are significant than high frequency coefficients. High frequency coefficients can be removed or reduced by quantization at different levels. Transform coding is lossy and complex compared to predictive coding. The details of compression methods used in WCE image compression are given below:

A Lossless and Near lossless Image Compression Algorithms

Lossless compression techniques reconstruct the compressed images without any loss in quality. Compression is achieved by exploiting the correlation between the pixels to remove pixel redundancy. In near lossless compression certain measure of quality loss is accepted without losing remarkable information. Compression is near-lossless when a small amount of loss is introduced due to quantization of the residue. Prediction based technique is a lossless compression algorithm, where the difference of a original and predicted pixel is encoded using Golumb Rice (GR) code. Complete pixel information is recovered at the decoder without any loss. These algorithms are very simple to implement and require very few computations for processing a pixel with low compression performance. JPEG-LS based compression algorithms provides better CR compared to other predictive techniques. The prediction of the pixel is done by the edge detection algorithm and the difference of the actual pixel and predicted pixel is encoded. The difference is encoded using fixed GR coding. When the flat regions are

Table 2.3: Predictive based WCE compression techniques

Study	Loss mode	Colour Space	Method
Simple prediction (Khan and Wahid (2011*a*))	Lossless	YUV	• Predicted value is computed by subtracting the neighbouring pixel. • Consists image clipping method and predicted values are entropy coded by GR encoding. • Compression efficiency is less than 60%.
DPCM+ Subsampling (Khan and Wahid (2011*b*))	Near Lossless	YUV	• Chroma components are sub-sampled and DPCM coded.
DPCM (Fante *et al.* (2016))	Both Lossy and Lossless	YUV	• Quantized and subsampled pixel values are DPCM coded to in lossy mode. • Pixels are DPCM coded without quantization and subsampling in lossless mode. • Adaptive GR entropy coding is employed.
Hybrid DPCM (Malathkar and Soni (2019))	Lossless	YEN	• An enhanced DPCM method • A modified signed Golomb code with bits skip code is used to improve the compression efficiency. • It enhances compression ratio by 2.3 % than conventional DPCM.
JPEG-LS (Chenb *et al.* (2009))	Near Lossless	RGB	• Compress using prediction and low pass filter before transformation. • Efficient compared to DPCM based techniques
JPEG-LS (Liu *et al.* (2016))	Near Lossless	RGB	• Before compression interpolation is done to improve the spatial correlation. • Due to interpolation compression efficiency reduces with reduced prediction error.

identified by the edge detection filter, the method uses run length encoding. The other type of most popular and low complexity predictive coding algorithms are differential pulse code modulation (DPCM) methods. Instead of just encoding the difference computed from previous element, these schemes use the prediction filter. These algorithms can work in lossless and near-lossless mode (Fante *et al.* (2016)). Lossless method encodes the residue without using the quantizer and near lossless methods uses quantizer. A summary of the existing lossless and lossy predictive methods is given in Table 2.3. Prediction based schemes are ideal for designing WCE image compressors due to its low complexity. Prediction schemes provide better quality but performs poor in terms of CR. Due to low compression efficiency, more data is left for transmission and consumes more power. This leads to fast draining of capsule battery power and causes incomplete scanning of the GI tract, when it is required to capture the images with high resolution and frame rate.

B Transform based Lossy Image Compression Methods

Lossy image compression methods use the discrete cosine transform (DCT) and discrete wavelet transform (DWT) to decorrelate images based on the correlation between pixels at the block level. These methods achieve the compression by exploiting spatial redundancy. Various transform based lossy compression methods used in WCE image compression is listed are Table 2.4.

DCT based methods: In DCT based compression methods (Lin and Dung (2011*a*), Turcza and Duplaga (2013)), the captured image from RGB colour space is converted to another colour space consisting of luma and chroma components using reversible colour space conversion. The chroma components are subsampled in the ratio 1:4 to achieve compression. These algorithms use 8 x 8 DCT on luma components and 4 x 4 DCT on chroma components. In the method proposed in (Gu *et al.* (2012)), 4x4 DCT is applied on R and subsampled GB components. In the work (Turcza and Duplaga (2017)), DCT is applied on every four pixels of the RGB components and difference of two DC coefficients is quantized and encoded to achieve near-lossless compression performance. In DCT based works, compression is achieved by quantizing the transformed coefficients which are later entropy coded to produce the compressed bit stream. The DCT transform results in higher compression performance, but the

Table 2.4: WCE lossy compression techniques

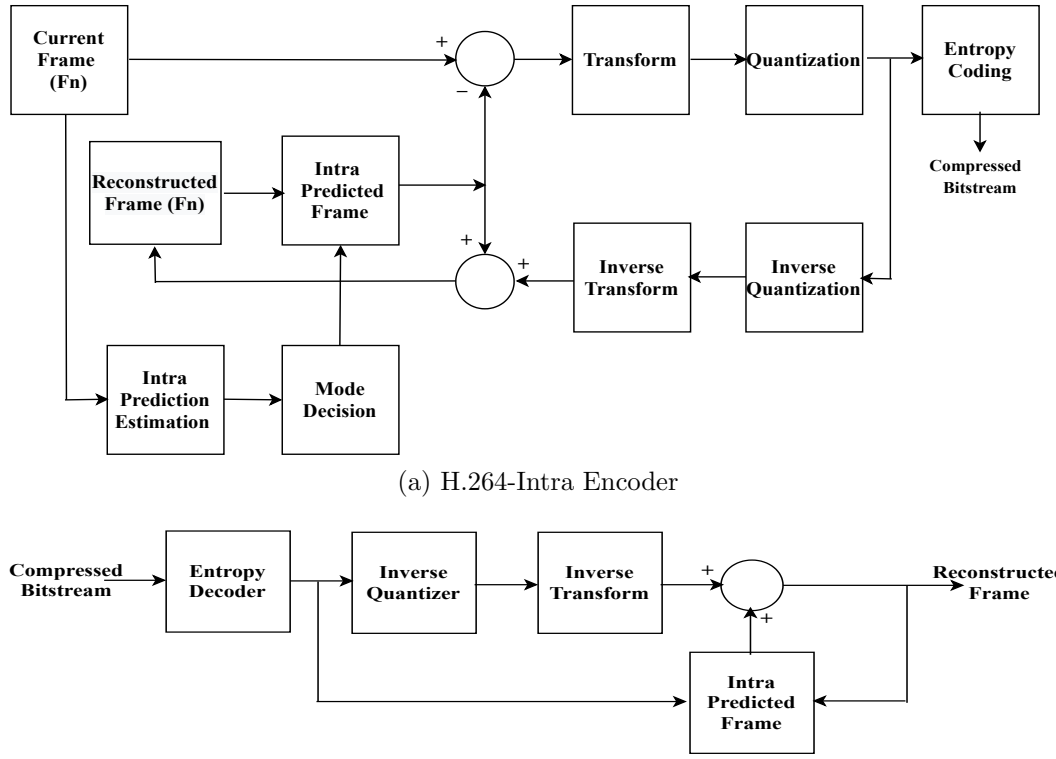
Study	Colour Space	Method
DCT (Lin and Dung (2011*a*))	YCoCg	• Before colour space transformation G and B components are subsampled at 2:1 and 4:1. • The transformed coefficients are quantized and entropy coded by Limpel-Ziv algorithm.
DCT (Turcza and Duplaga (2011))	YCoCg	• Transformed coefficients are DPCM coded along with variable Huffman length coding.
DCT (Gu *et al.* (2012))	RGB	• G and B components are subsampled at 2:1. Transformed and quantized coefficients are Huffman coded.
DCT (Turcza and Duplaga (2017))	RGB	• DCT with predictive coding with near lossless coding. • Transformed coefficients are GR encoded.
DWT (Thoné *et al.* (2010))	YCbCr	• Haar wavelet transform is used. • Run length Huffman encoding is used for entropy coding. • Achieves high compression efficiency compared to DCT based techniques. • High computational complexity and memory requirement.
Modified H.264-Intra (Dung *et al.* (2008))	RGB	• Only DC intra prediction mode is used.

amount of information loss in the reconstructed images is significant. DWT does not have this limitation and can be used to achieve high compression with high quality.

DWT based method: WCE compression method proposed in (Thoné *et al.* (2010)) is based on DWT which uses an analysis filter bank, where the image is passed through a series of lowpass and highpass operations. This process decomposes the image into sub-images which are further considered for quanitization and compression. Though the DWT based method achieves better compression at higher PSNR compared to DCT based transform methods, it consumes more memory on the encoder as the entire frame needs to be stored for performing analysis filter bank operations. Also more computations are required for processing the frame and consumes more power and area.

H.264-Intra method: H.264-Intra is popularly used for medical image compres-

sion which requires high quality reconstruction. However, because it uses complex rate distortion optimization techniques at the encoder, it is more difficult to apply in WCE image compression, even though it gives superior compression performance than other WCE image compression techniques with good quality metrics. Therefore, H.264-Intra method modified to operate only in 4 x 4 DC prediction mode is proposed to reduce the complexity (Dung *et al.* (2008)). Detailed functional description of H.264-Intra is given in this section as it is considered for comparing the performance and complexity of the methods presented in Chapters 3 and 4.



(a) H.264-Intra Encoder

(b) H.264-Intra Decoder

Figure 2.6: H.264-Intra frame coding system

The block diagram of the H.264-Intra frame coding system is shown in Figure 2.6. H.264-Intra encoder consists an encoding path and reconstruction path. Encoding path consists transform, quantization and entropy coding blocks which takes an input frame and creates a compressed bitstream. The encoded frame is decoded and reconstructed using the reconstruction path to guarantee that both the encoder and decoder gets the similar reference frame for intra prediction. This eliminates the possibility of encoder-decoder mismatches as the decoder never receives the original

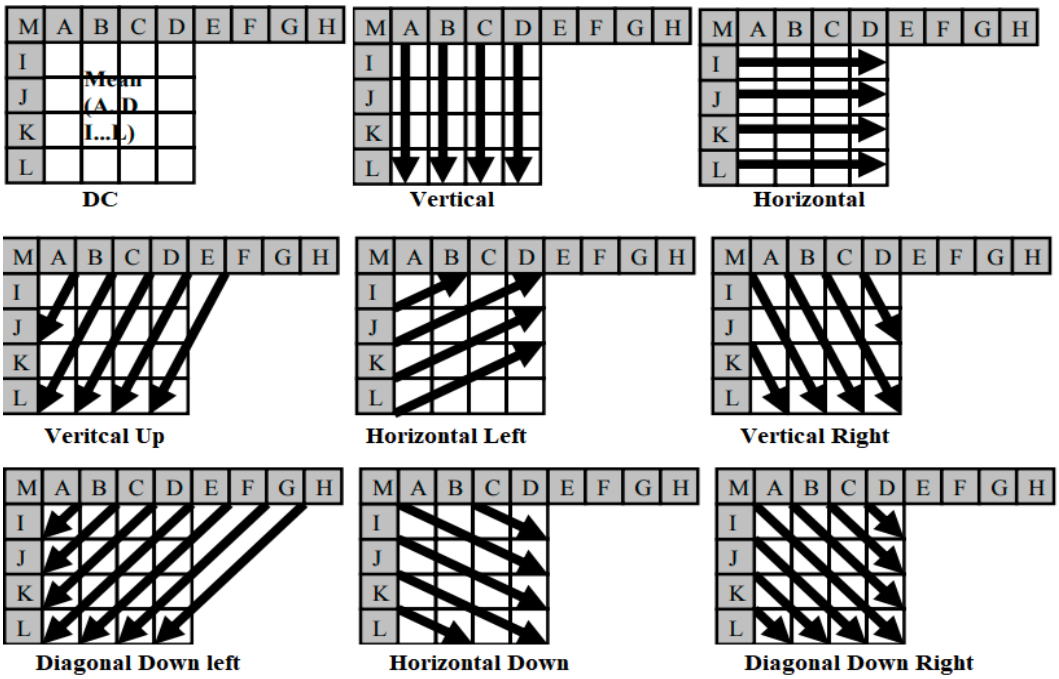


Figure 2.7: Intra prediction modes for 4x4 luma block

frame. For an input frame in YCbCr 4:2:0 sampling format, the luma (Y) component is divided into macroblocks (MB) of 16x16 pixels and chroma (Cb,Cr) components are divided into blocks of 8x8 pixels. Each macroblock is encoded by using different intra-frame prediction modes. Each 4x4 luma block has nine directional prediction modes (Hamzaoglu *et al.* (2008)) and each mode predicts 16 pixel values from neighbouring block pixels A to M as shown in Figure 2.7. Pixels A to M are considered to be encoded and reconstructed previously and the same are available for the encoder and decoder to generate the prediction. A 16x16 MB has 4 prediction modes normally selected for the smooth regions while 4x4 prediction modes are preferred for textured regions. The 8x8 MB of chroma component consists 4 prediction modes. The residual frame is generated by subtracting the predicted frame from the current frame and transformed using integer DCT. The transformed coefficients are quantized and entropy coded to generate compressed bitstream using context adaptive variable length coding algorithm. To reconstruct the residual frame, the quantized components are inverse quantized, inverse transformed and added to the predicted frame to reconstruct the current frame.

Limitations: The transform coding methods provides high compression performance compared to predictive techniques with acceptable image quality. But these

methods cannot exploit inter frame correlation to remove temporal redundancy which is high in WCE video content. This will result in low performance in terms of compression. Low complexity video compression techniques can be employed to remove the inter frame redundancy to provide better compression performance as explained in Section 2.3.3.

2.3.3 WCE Video Compression Methods

Better compression can be achieved by removing the spatial and temporal redundancy which results due to strong correlation between the block of pixels within the frame and neighbouring frames respectively. The compression system which removes the temporal redundancy by considering group of consecutive frames is called as video coding system. The system consists an encoder to compress the video and the decoder to decompress the encoded video to reconstruct the original video. The compression performance is assessed using bitrate measured in bits per second (bps) and the quality. PSNR and SSIM metrics are used to assess the quality of the reconstructed video.

In the WCE video, consecutive frames are highly correlated due to the slow movement of the capsule. While travelling through the small intestine, the capsule exhibits slow motion and sometimes no motion. Around 50% of the total frames are captured in the small intestine are highly correlated in terms of time. The frames captured in the other GI tract organs have a moderate temporal correlation. The temporal redundancy can be reduced by using video compression techniques to achieve higher compression performance. In addition to this, it is apparent that capsule endoscopes experience major limitations in terms of processing capability and power consumption. Therefore, low complexity compression algorithms must be utilised to avoid the capsule being bulky and the image compressor itself consuming additional power. Furthermore, as the recorded video is used for medical diagnostics, providing high-quality decoded video with a high compression ratio is essential.

Standard video coding systems: Many standard video compression schemes such as MPEG standards, H.264/AVC (Wiegand *et al.* (2003)) and H.265/HEVC (Sullivan *et al.* (2012)) exist that focus on obtaining optimized compression performance by exploiting inter frame correlation. These standard video coding architectures com-

press the video many times by removing the redundancy at the encoder by employing computationally expensive motion estimation task. Hence, the standard video compression architectures with full inter coding are unsuitable for attaining efficient video compression at low power consumption with enhanced battery life requirements set by the capsule endoscope. Motion JPEG (MJPEG) is another standard video coding system popularly used to compress medical video. MJPEG encodes each frame of a video sequence separately in JPEG format. It is computationally less complex but performs poor in terms of compression performance as it does not remove the temporal redundancies between the successive frames. The main blocks involved in MJPEG encoding are colour-space conversion, chroma sub-sampling, DCT, quantization and entropy encoding.

DVC: DVC focuses on developing a low complexity encoder for power and resource constrained devices by shifting complex motion estimation from the encoder side to the decoder side. On the other hand, there is no restriction for resource and power consumption on the decoder. Therefore, DVC is more suitable for the applications such as WCE where the video is captured and encoded by a power constrained device and decoded by the powerful computer without any time restriction. DVC is framed based on two well-known theorems put forward by Slepian-Wolf known as SW coding (Slepian and Wolf (1973)) and its extended version Wyner-Ziv coding popularly called WZ coding (Wyner and Ziv (1976)).

TDWZ-DVC: Transform domain Wyner-Ziv based DVC (TDWZ-DVC) architecture shown in Figure 2.8 gives better compression performance compared to other DVC systems and it is the preferred architecture in most of the research. The functional description of TDWZ-DVC is given in this section as it is considered for comparing the performance and complexity of the proposed WCE video compression methods.

In TDWZ-DVC, an input video sequence is split into group of pictures (GOP) and each GOP consists of the initial keyframe and remaining WZ frames. GOP format of size 2,4 and 8 is shown in Figure 2.9. Each frame consists only luma component and coding of chroma components is not addressed in this codec. The keyframes are encoded by H.264 codec in intra mode and WZ frames are Wyner-Ziv coded. WZ coding of each frame involves transform coding of blocks, quantization and bit-plane

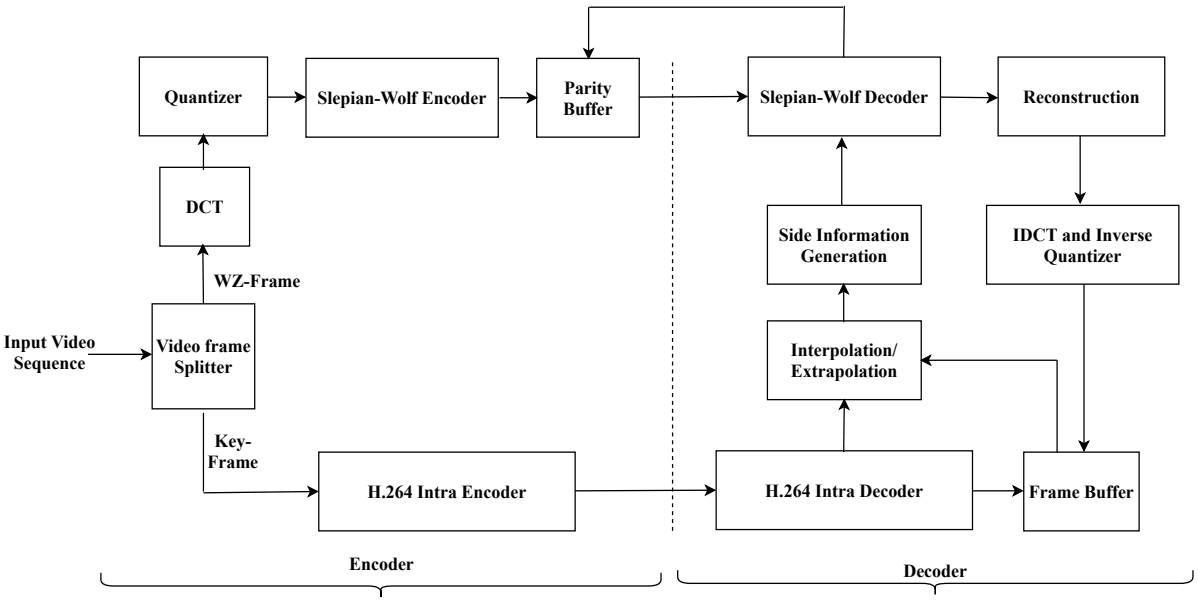


Figure 2.8: TDWZ-DVC architecture

Frame Num	1	2	3	4	5	6	7	8	9	10	11	12	.....
GOP-2	Key	WZ	Key	WZ	Key	WZ	Key	WZ	Key	WZ	Key	WZ	.....
GOP-4	Key	WZ	WZ	WZ	Key	WZ	WZ	WZ	Key	WZ	WZ	WZ	.....
GOP-8	Key	WZ	WZ	WZ	WZ	WZ	WZ	WZ	Key	WZ	WZ	WZ	.....

Figure 2.9: GOP formats in DVC

formation. The WZ frames are transformed using 4x4 DCT and the transformed coefficients are uniformly quantized. The coefficients are then grouped into separate frequency bands, with each band containing the same frequency coefficients in different blocks. The bits for each band are extracted and organised into bitplanes. These bitplanes are encoded to compute parity bits by systematic channel encoder, generally a low density parity check (LDPC) accumulate. The parity bits generated by LDPC coding of the bit-planes are stored in the parity buffer and transmitted whenever the decoder sends a request signal. The amount of parity bits transmitted depends on the quality of side-information (SI) generated at the decoder using previously reconstructed frames. SI is considered as the current WZ frame transmitted with errors. Parity bits are used to correct the errors at the decoder. The decoder will stop sending the request signal when the LDPC decoder corrects all the errors. This process is called as Slepian-Wolf coding.

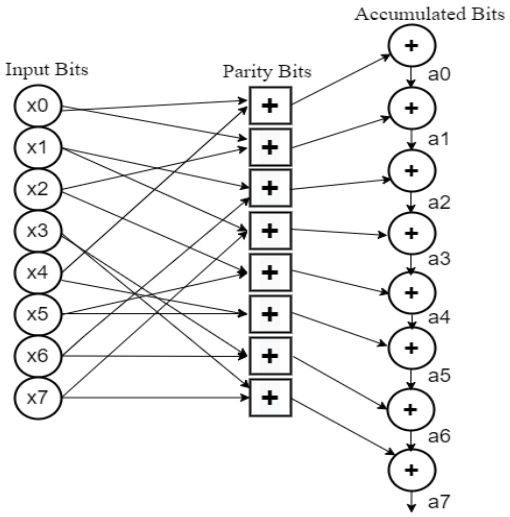


Figure 2.10: LDPCA Encoder

The LDPCA encoder shown in Figure 2.10 is the combination of parity bits generator and an accumulator. The bits in each bitplane are EX-ORed at the parity nodes using the LDPC graph structure constructed using the method proposed in (Varodayan *et al.* (2011)) to produce parity or syndrome bits. The accumulated parity bits are produced by EX-ORing parity bits and are buffered in the parity buffer. The buffered bits are transmitted in chunks when the request signal is received from the decoder.

The side information for every WZ frame is generated by the decoder through motion compensation of frame interpolation or extrapolation of the previously decoded closest frames. The bitplanes formed by DCT coefficients of the side information are LDPCA decoded. The LDPCA decoder modifies its structure when an additional chunk of the accumulated parity bits are received. Syndrome bits are retrieved on the decoder side by EX-ORing the consecutive accumulated parity bits and used in refining side information. The correctness of the side information refinement is tested by using syndrome bits. The performance of the DVC system relies on the quality of side information generation at the decoder. Better quality side information need few parity bits for the refinement and achieves better compression, where as poor quality SI need more bits for the correction and degrades the performance. After LDPCA decoding, the decoded quantized symbol streams are formed by combining the bitplanes of each DCT band. Next, DCT coefficients are reconstructed when all the quantized symbols are available. The final WZ frame reconstruction is completed

by inverse quantization and Inverse DCT and the reconstructed WZ frame is saved in the frame buffer.

DVC techniques proposed for non-WCE applications use interpolation of previously decoded frames to generate SI. This introduces buffering complexity at the encoder due to the storage of WZ frame bit-planes and restricts the GOP size to 2. The limitation on GOP size increases the number of keyframes for encoding. As a result, the use of temporal correlation in additional frames will be limited, resulting in poor compression performance. WCE frames exhibit irregular motion, and interpolation techniques cannot produce better SI. To overcome the issue of SI generation for WCE video, hash driven DVC is proposed in (Deligiannis *et al.* (2011)). Downsampled version of WZ frame is used as an hash which is intra coded and transmitted. At the decoder, block based motion estimation technique is used to generate SI. But in this method SI generation takes more time and introduces latency in decoding, which limits the frame rate. Moreover, hash creation and transmission at the encoder is an extra overhead when low complexity is desirable. In addition to all these, using H.264 Intra for keyframe encoding increases the complexity of the encoder.

Adapted vector quantization (AVQ) based SI creation for DVC with a highly computationally intensive searching method is presented for WCE in (Boudechiche *et al.* (2017)). Codebook consisting of WCE frames of the entire video is used for SI creation. Since the code book is used for the SI creation, all the frames are treated as WZ frames eliminating the keyframe encoding. Although this method reduces the cost of encoding keyframes to a large extent, using the available image database to create the SI is not a superior option because capsule motion is unpredictable and varies from person to person. Another disadvantage of this method is that it results in a poor frame rate due to the increased complexity of the decoder search. This adds to the buffering complexity by causing a greater delay in SI creation. Delay complicates the buffering of parity bits and slows the frame transmission rate.

2.3.4 Compression Algorithms in Commercially Available Capsules

There are many commercial capsules manufactured by three leading companies in the field of GI endoscopy such as Given Imaging Inc., Intromedic, Olympus America

which are approved by food and drug administration (FDA). Details and techniques used in designing compression systems for commercial capsules are not available in publications. Some abstract information collected from the patents which are publicly available is provided in Table. 2.5.

Table 2.5: Compression methods used in commercially available capsules

Manufacturer & Capsule Name	Patent Inventors	Compression method
Medtronic Pillcam	*Glukhovsky et.al.* (Glukhovsky *et al.* (2003))	• Compression methods based on JPEG and MPEG standards.
	Zinaty et.al. (Zinaty *et al.* (2015))	• Each image pixel is transformed into another colour space (Y, Cb,Cr,Gdiff) before compression. • Low complexity fast lossless compressions method is used.
	Avni et.al. (Avni *et al.* (2010))	• The difference of two image pixels is encoded and transmitted rather than the original image.
	Horn et.al. (Horn (2008))	• Images are acquired with variable frame rate depending on organ of interest in GI tract. • Images are compressed when they are captured at higher fps.
Olympus America Endocapsule	*Shigemori et.al* (Shigemori and Matsui (2011))	• JPEG based compression technique is used with the removal of some pixels.
	Bandy et.al. (Bandy *et al.* (2013))	• Images are captured with variable frame rate and resolution with respect to speed. • Low resolution and low frame rate is used at lower speed and image is compressed by grouping blocks of pixels. • Higher resolution and high frame rate is used when the speed is high. • This concept ensures that most of GI tract is fully scanned to minimize the missing rate of lesions. • A new concept is used where frame rate and resolution varies based on the speed of the capsule.

2.3.5 Research Gap in WCE Video Compression

TDWZ-DVC provides a low complexity encoding solution for an application such as WCE. The performance of the TDWZ-DVC system rely on the quality of SI generation at the decoder. For SI generation most of the DVC based architectures use motion

compensated interpolation and extrapolation of previously decoded frames (Borchert *et al.* (2007), Brites and Pereira (2008)). An hash is used in DVC to improve SI of WCE (Deligiannis *et al.* (2011)). Hash is generated using down-sampling the WZ frame which is intra coded and at the decoder overlapped block based motion estimation (OBME) is used to generate better SI. Another DVC method based on adapted vector quantization (VQ) with a highly complex searching method is used for WCE (Boudechiche *et al.* (2017)). VQ allows for the creation of SI from a codebook instead of motion compensated prediction of keyframes. Keyframe encoding is eliminated and all the frames are treated as WZ frames. But it uses an available database for the creation of SI which might not be a suitable solution as capsule movement is irregular and varies from patient to patient. Also, high searching complexity at the decoder may cause more delay which increases the parity bits buffering complexity and limits the frame transmission rate.

The existing TDWZ-DVC architectures cannot be used directly for compressing WCE video. Some modifications are required to be incorporated to make DVC suitable for WCE video compression due to the following:

- Keyframes are encoded by H.264 intra-frame encoder. H.264 intra prediction employs a RD optimization method which increases the encoding computational complexity significantly. The computational requirements for intra prediction modes is too high for an application such as WCE. The complexity of the keyframe encoder should be as low as JPEG compression algorithm which has color space conversion, sub-sampling, DCT, coefficient quantization and entropy coding.

- Keyframes are encoded by conventional intra coding methods do not exploit the textural characteristics of WCE images to achieve better compression.

- Decoding of WZ frames is not possible unless past and future keyframes are decoded. Till the next keyframe is encoded and decoded, WZ frame is buffered which introduces latency and causes buffering complexity at the encoder. To avoid buffering of more number of frames, a GOP is limited to two which increases the number of keyframes and decreases the performance of DVC.

- OBME method of SI generation takes a long time and induces latency in decoding. Also, the hash information is an extra overhead, though it is transmitted at a low resolution at a very low quality and formation of the hash at the encoder increases the complexity.

- Consecutive frames of WCE video captured in a particular GI organ has very little motion or no motion, or sometimes large motion. Irrespective of the motion characteristics, a small region in a GI tract exhibits homogeneity in colour and texture (Khan *et al.* (2015)). Therefore it is sufficient to transmit only the luma of WZ frame and chroma components of WZ frame can be generated at the decoder side by matching luma and texture information of neighbourhood keyframes. The elimination of WZ-chroma processing and transmission saves both in terms of processing time and complexity, which is indispensable for power conservation in battery operated capsule.

2.4 Literature Review on WCE Video Summarization

The problem of video summarization (VS) can be described as selecting a small batch of frames from the video stream consisting of a large set of video frames that describe the whole content of the original video. VS is a technique for parsing the sequence of video frames into a shot set and extracting the set of keyframes (Zhu *et al.* (2005)). Most of the state-of-the-art VS methods use three main common steps (Gygli *et al.* (2014)):

- Feature extraction from each frame and latent space representation of extracted features.
- Temporal segmentation of video into shots. Each shot consists a group of sequential frames with certain similar features. Frames in each shot consist of similar kinds of frames with respect to colour, shape and texture with small motion.
- Finally, a set of frames called keyframes are extracted from each shot which describes the entire content of the shot.

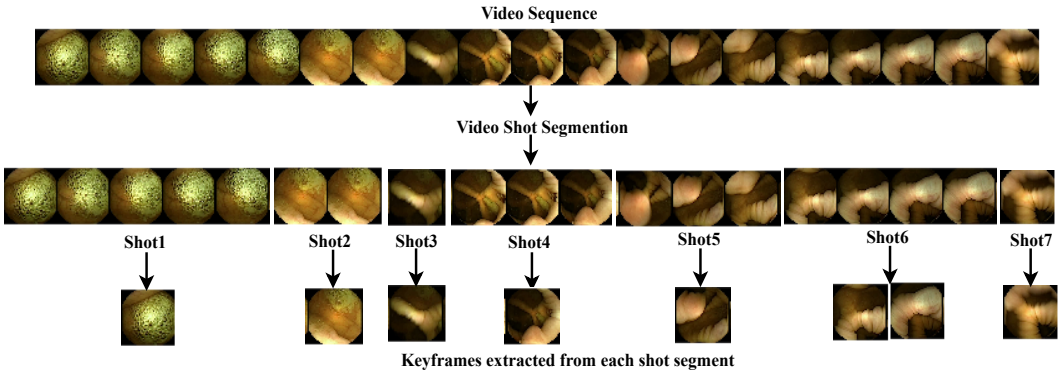


Figure 2.11: Typical keyframe extraction from a video sequence

A typical video summarization method to extract keyframes is shown in Figure 2.11. The details of the existing WCE video summarization techniques are described in the following sections.

2.4.1 Summarization based on Handcrafted Features Extraction

In most of the methods proposed for WCE video summarization, colour, shape and texture features are used in finding similarity between frames. In computer-aided imaging techniques, features such as colour, texture and shape are considered as handcrafted features representing low level features. These features are acquired by using various algorithms from the information content of an image itself. Primarily these features are used in conventional machine learning based computer vision applications such as image classification and recognition. The information from the handcrafted features is used for similarity estimation between the frames in WCE video summarization.

In work (Mehmood *et al.* (2014)), features such as image moments fusion, multi-scale contrast and curvature are used to calculate the visual saliency map for every frame. Image statistical moments such as mean, skewness, standard deviation and kurtosis are computed to describe the structural shape, boundaries and textural features in an image. These features are utilized to compute the similarity between two consecutive frames. Multi-scale contrast for every pixel in a frame produces the grey level saliency map. The normalized fusion of image moments, contrast and curvature measures is used to find the final saliency map. The resultant saliency map identifies the keyframes by using the threshold.

Table 2.6: WCE summarization techniques using Hand-crafted features

Study	Features type	Feature extraction techniques	Shot detection method	Keyframe extraction
Li *et al.* (2010)	Colour	Colour histograms	Distance	K-means clustering
Zhao and Meng (2011)	Colour, texture and shape	color moment invariants, LBP and Fourier coefficients	Euclidean distance between features	Linear discriminant analysis
Chen *et al.* (2012)	Edges	Canny detection	No shot detection	Euclidean distance between edge shift
Huo *et al.* (2012)	Colour and texture	HSV and Block edge directivity descriptor	Frame texture difference	Relational rank matrix
Yuan and Meng (2013)	Colour, texture and shape	HSV, LBP and HoG	Information entropy	Affinity propagation clustering
Ismail *et al.* (2013)	Colour and texture	HSV and Edge Histogram Descriptor	Fuzzy-C means Clustering	Only outliers are eliminated
Lee *et al.* (2013)	Intensity	Normalized cross correlation	Frame similarity	Motion analysis with SURF matching
Liu *et al.* (2013)	Motion features	Bee algorithm SIFT descriptor method	Changes in scene measurement	Forward and backward motion estimation
Mehmood *et al.* (2014)	Shape, boundaries and texture	Mean, skewness, standard deviation and kurtosis	Fusion of Image moments, contrast and curvature	Saliency map thresholding
Chen *et al.* (2015)	Colour and texture	HSV and Gray level co-occurrence matrix (GLCM)	Frame similarity measure and adaptive threshold	Adaptive K-means clustering

Edges are utilized as features in the frames to quantify distance, and the threshold is used to determine frame similarity. When the edge shift between two frames is less than a predetermined threshold, the second frame is considered redundant (Chen *et al.* (2012)). Multiple features such as colour, texture and shape are extracted to detect the shot and from each shot the keyframes are extracted using a linear discriminant analysis algorithm (Zhao and Meng (2011)). In another work proposed in (Yuan and Meng (2013)), colour features are extracted by Hue, saturation and value (HSV) and shape features by using histogram of oriented gradients (HoG). For each feature information, entropy is computed to find the shot. From each shot, the keyframes are extracted by using affinity propagation clustering algorithm.

In work (Guo *et al.* (2010)), mainly the colour features, texture features extracted by local binary pattern (LBP) and shape features are used. A fixed threshold is used to segment the video into different shots based on the distance between the features. However, a fixed threshold is not a desirable choice, and hence other works go for the segmentation with an adaptive threshold (Chen *et al.* (2015)). Keyframes are the most representative frames in a video that may be retrieved using machine learning techniques including linear discriminant analysis, relational rank matrix, K-means clustering, and speeded up robust features (SURF) (Bay *et al.* (2008)). Various methods along with the performance based on feature based video summary generation techniques are listed in Table 2.6.

2.4.2 Non-Matrix Factorization based Unsupervised Methods

The Non matrix factorization (NMF) methods are applied on the clusters to extract the most representative frames of the video sequence. The input video sequence is divided into clusters based on the similarity estimation between the frames. NMF retains much of the structural details of the input video frames by using non-negative basis and its related weights. NMF algorithm takes a matrix $m \times n$ as input and tries to find W and H. W represents the matrix consisting of the NMF basis and H is a matrix of non-negative weights. NMF uses low-dimensional subspace to provide frame reduction by offering the most representative frames that cover the entire GI examination video.

NMF based frame reduction method, which involves three steps is proposed in (Tsevas *et al.* (2008)). Initially, the feature dimensionality reduction technique is used to get the matrix of non-negative values. Fuzzy-C-means clustering method is used to cluster the frames into a predefined number of groups in the second step. Finally, the NMF algorithm is used to extract the keyframes. In another clustering based method proposed in (Iakovidis *et al.* (2010)), the keyframes are extracted by adaptive based threshold instead of using a fixed threshold from each cluster to avoid losing the significant frames.

2.4.3 Deep CNN based Learning Techniques

Deep learning methods are used to extract the high level features to overcome the limitations of methods based on handcrafted features. A deep convolutional neural network (DCNN) can extract the high level features required to classify the consecutive pair of frames as similar and dissimilar pair of images. A video shot boundary is detected based on the similarity between the frames. From each shot, keyframes are extracted to remove the redundant frames.

Siamese neural network (SNN) is trained in a supervised manner to extract the features in work proposed in (Chen *et al.* (2016)). Support vector machine is trained using a small set of images to classify the images as similar or dissimilar. The method proposed in (Biniaz *et al.* (2020)) used a pre-trained DCNN model to extract the high level features. The radial basis function is used to construct the high dimensional model from the high level and low-level feature space. A singular value decomposition based adaptive sliding window method is used to extract the keyframes.

2.4.4 Research Gap in WCE Video Summarization

In the WCE video, colour and texture content varies slightly from one frame to the next consecutive frame. Therefore, colour and texture features are insufficient to detect significant changes between two successive frames (Primus *et al.* (2013)), resulting in an inaccurate video summary and the possibility of missing important frames with significant lesions. This has a lot of practical implications where accuracy is a primary requirement in medical diagnosis.

Deep learning based summarization approach perform better compared to conventional methods that depend on weighted fusion of handcrafted features and frame clustering algorithms (Apostolidis *et al.* (2021)). In deep learning approach, CNNs are used to extract the high level features from the frames of the input video sequence. Large set of videos and groundtruth summaries are required to train deep CNNs in a supervised way. Creating groundtruth summaries for WCE video is a time consuming and tedious task that also necessitates the assistance of a gastroenterologist. Hence, there is a need of an unsupervised method to train deep CNN to generate WCE video summary.

Chapter 3

Distributed Video Coding Architecture with Frequency Band Classification

3.1 Introduction

The low complexity encoder requirement and energy constraint problem of WCE can be potentially solved by implementing DVC. This chapter presents a DVC architecture for WCE video encoding that attempts to overcome the drawbacks discussed in Section 2.3.5. The video is encoded in DVC as a group of pictures (GOP), with an initial keyframe and the subsequent WZ frames encoded using two different techniques. A JPEG based keyframe encoder with modifications at the transform and quantization stages is proposed as a replacement of H.264-Intra to reduce the complexity. To further reduce the number of computations at the quantization and entropy coding levels, the proposed keyframe encoder adopts WCE image textural characteristics. A simple technique is proposed for hash generation to improve the quality of SI. The proposed method also presents a new approach for encoding subsampled chroma components where in the SI generation of chroma components is done by using the low bands of the luma component without using separate hash explicitly for chroma. SI generation in the proposed system depends only on the previously decoded frames which removes the restriction on GOP size and hence improves the compression performance and reduces the encoder buffering complexity. Latency in SI generation is also reduced as SI generation depends only on previously decoded frames.

3.2 DVC Architecture for WCE Video Coding

The DVC architecture suitable for WCE video coding is shown in Figure 3.1. The images captured by WCE image sensor in RGB colour space are transformed into YC_bC_r colour space. Chroma (C_bC_r) components are subsampled using YC_bC_r 4:2:0 subsampling format. Next, the frames in each GOP of the video sequence are split into keyframe or WZ frame. The JPEG based encoder encodes the keyframe by adopting WCE image textural properties to classify the image blocks into smooth and non-smooth blocks in the transform domain to reduce the computational complexity. The detailed explanation of the proposed keyframe encoding method is given in Section 3.3.

In most of the low complexity image and video compression applications, integer DCT is widely used due to its good energy compaction properties. But, for the images with smooth textural properties such as GI tract images, integer discrete tchebichef transform (DTT) performs better than integer DCT. The implementation of integer DTT needs only addition and shift operations which reduces the computational complexity compared to integer DCT. Therefore in the proposed system, DTT is used to transform an image block from spatial to frequency domain and its significance on compressing WCE images is given in Section 3.3.1. Modified JPEG quantization table is used for quantizing the transformed coefficients and need only bit-shift operations. In the proposed system, 8x8 transform is used for luma to reduce the number of intra bands and bitplane length.

In WZ frame encoding, the three low frequency quantized luma coefficients of each block are intra coded. Remaining high frequency components are considered for forming higher frequency bands. The subsampled chroma components are transformed using 4x4 DTT. The quantized higher frequency bands of the luma and lower 6 frequency bands of chroma components of the WZ frame are WZ encoded. The quality of SI generated at the decoder has a significant impact on WZ encoding performance. If the quality of SI estimation is not good, WZ coding performs worse than intra coding. Therefore, the proposed system encodes few luma components in intra mode using adaptive Golomb Rice (AGR) encoder which is used as an hash to create better SI at the decoder and the remaining components in WZ mode. The same hash created using luma components is used to generate SI for chroma. Detailed explanation

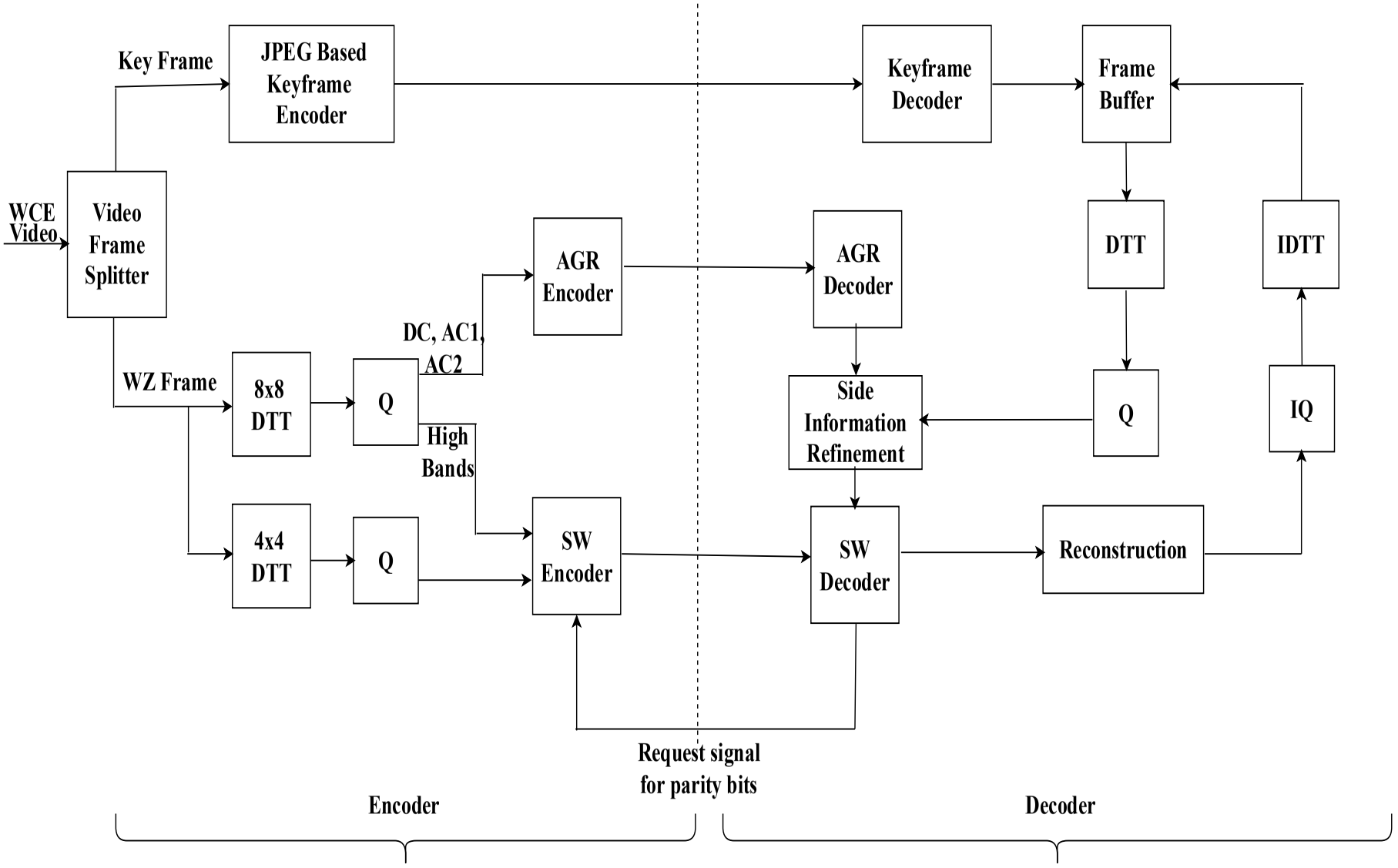


Figure 3.1: DVC based architecture with WZ frame coding based on frequency band classification. Block Q is Quantizer, IQ is inverse Quantizer and SW refers to Slepian-Wolf coding.

for WZ coding of luma and chroma components is given in Sections 3.4.1 and 3.4.2. The method used for SI generation and refinement is given in Section 3.4.3. The proposed system is referred as DVC with frequency band classification (DVC-FBC) in the remainder of the thesis.

3.3 Keyframe Encoder

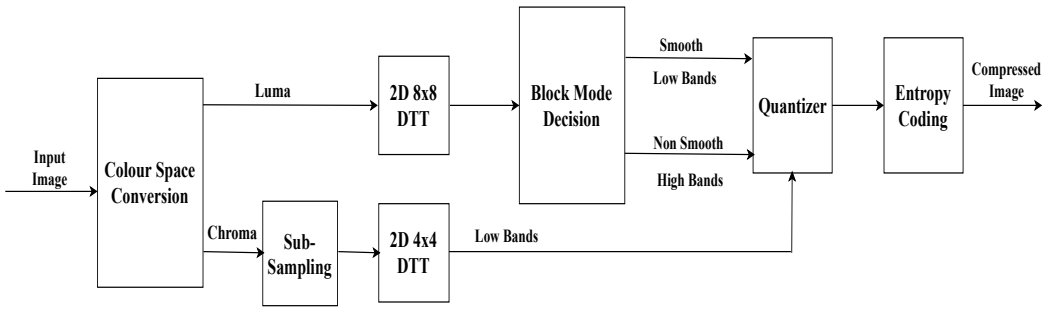


Figure 3.2: JPEG based key-frame encoder with block mode decision

Keyframe encoder considers the WCE image texture characteristics to get better compression performance at reduced computational cost. Block diagram of keyframe encoder is shown in Figure 3.2. It mainly consists of colour space transformation, subsampling of chroma, image transformation into frequency domain, block classification, quantization and entropy coding. These functional blocks are explained in the following sections. In the remaining of the chapter the proposed keyframe encoder is referred as block texture conditioned keyframe encoder (BT-KFE).

3.3.1 Image Transformation using Approximate DTT

WCE images are highly correlated in the spatial domain. On a pixel block of size $N \times N$, a linear orthogonal transformation is used to reduce the strong spatial correlation and provide energy compaction into very few coefficients. A linear transformation $T : R^n \rightarrow R^n$ is said to be orthogonal for all $x, y \in R^n$, if it satisfies $\langle T(x), T(y) \rangle = \langle x, y \rangle$. R^n is real inner product space, that preserves the inner product between x and y after the transformation. A 2D 8x8 transform is used for luma channel and 2D 4x4 transform is used for chroma channel. In terms of average bitlength and quality, DTT performs better than DCT as reported in (Prattipati *et al.* (2013)) for the images with

smooth texture. The 2D DTT for a 2D input sequence of order NxN is defined as,

$$Y(k_1, k_2) = \sum_{n_1=0}^{N-1} \sum_{n_2=0}^{N-1} \Lambda(k_1, n_1)\Lambda(k_2, n_2)x(n_1, n_2) \tag{3.1}$$

where $k_1, k_2 = 0, 1, \ldots N-1$ and $\Lambda(k, n)$ represents the orthogonal basis of DTT and is given as,

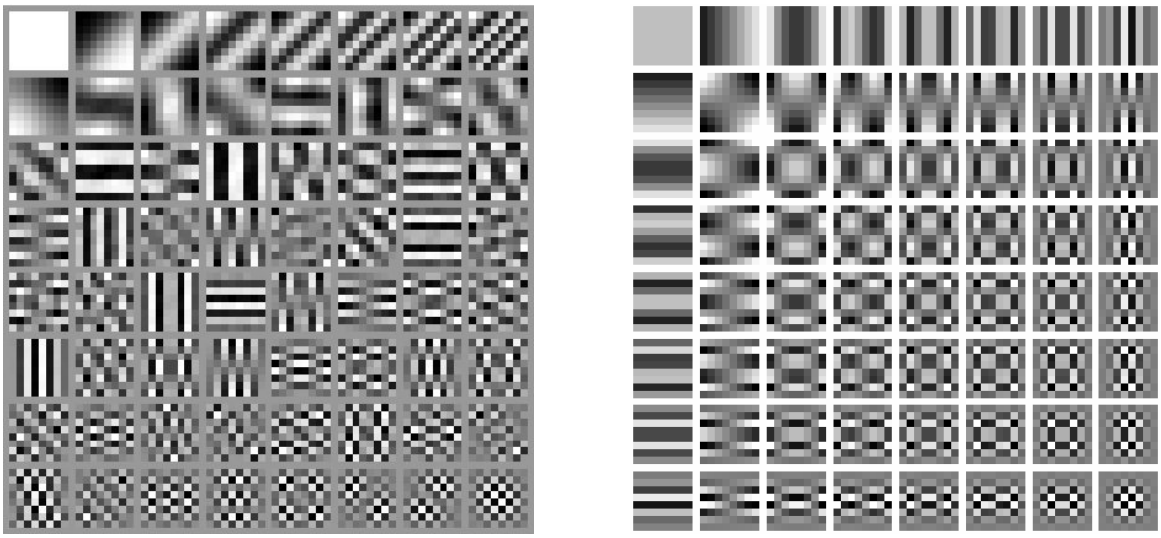
$$\Lambda(k, n) = (a_1 n + a_2)\,\Lambda(k-1, n) + a_3\Lambda(k-2, n) \tag{3.2}$$

with

$$\Lambda(0, n) = \frac{1}{\sqrt{N}}, \;\; \Lambda(1, n) = (2n + 1 - N)\sqrt{\frac{3}{N^3 - N}}$$

$$a_1 = \frac{2}{3}\sqrt{\frac{4k^2 - 1}{N^2 - k^2}}, \;\; a_2 = \frac{1 - N}{k}\sqrt{\frac{4k^2 - 1}{N^2 - k^2}},$$

$$a_3 = \left(\frac{1 - k}{k}\right)\left(\frac{2k + 1}{2k - 3}\right)\sqrt{\frac{N^2 - k^2 + 2k - 1}{N^2 - k^2}}.$$

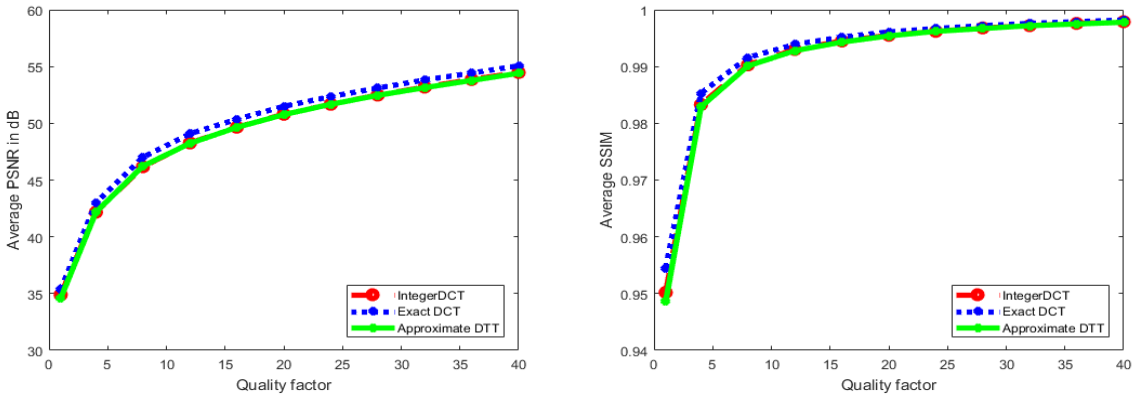


(a) (b)

Figure 3.3: Basis images of 8x8 (a) DCT and (b) DTT

Generally, fast integer DCT which is the approximated version of floating point also called as Exact DCT (Huang *et al.* (2019)) is used in many compression algorithms to reduce the computational cost compared to other versions of DCT. In this work, multiplier less approximate integer DTT (Oliveira *et al.* (2017)) is used as it requires only 384 additions and 96 bit-shift operations to compute 2D 8x8 transformation compared to integer DCT which needs 512 additions and 224 bit-shift operations (Gordon

et al. (2004)). The complete set of basis images for the DCT and DTT are shown in Figure 3.3. Average PSNR and Average SSIM have been evaluated for the available data for both integer DCT and approximate DTT.



(a) (b)

Figure 3.4: (a) Average PSNR and (b) Average SSIM measurements of WCE images with quantization at different quality factor for the considered transforms

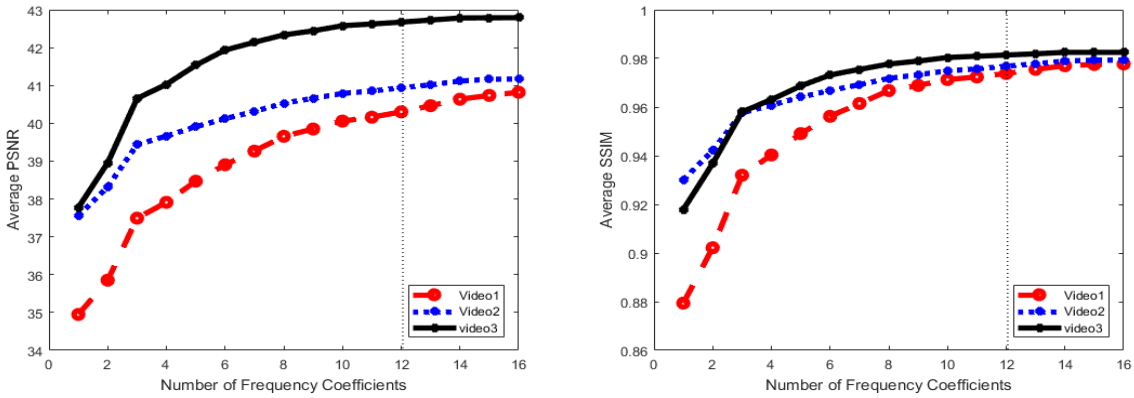
PSNR and SSIM for around 300 endoscopic images are computed and an average of that is considered for evaluating performance of considered transforms. Average PSNR and SSIM measurements for WCE images for the transforms considered at different quality factor (QF) is displayed in Figure 3.4. It is observed that the performance of approximate DTT at reduced computational cost for endoscopic images is the same as integer DCT.

3.3.2 Smooth and Non-smooth Block Mode Decision

The healthy GI tract has smooth and uniform coloured tissues. The pixel distribution is consistent in a smooth region. When the blocks in smooth region are represented in the frequency domain, significant energy is present only in low frequency components while the high frequency components become zero after quantization. If the block has edges and is textured, it is treated as non-smooth block resulting in significant high frequency components. Usually, the abnormal tissue regions exhibit textured pattern or otherwise endoscopic tissues are smooth in nature. In this method, the blocks are classified into smooth and non-smooth blocks before quantization based on energy content in lower frequency bands.

Smooth block represented in the frequency domain has more energy compaction in DC component compared to AC components. To detect the block mode of an 8x8 block, magnitude of DC coefficient is compared with the sum of magnitudes of first two AC coefficients in the first column, first two AC coefficients in the first row and first two AC coefficients in the diagonal to DC value. Block mode is viewed as smooth if the energy in DC coefficient is around ten times greater than the sum otherwise non-smooth. Chroma blocks are considered as smooth blocks as they do not exhibit much variation in adjacent pixels.

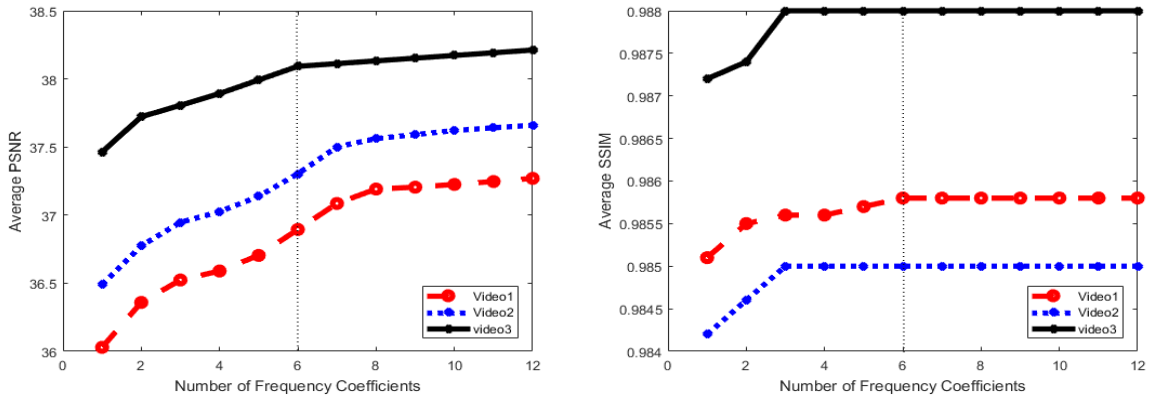
3.3.3 Quantization



(a) (b)

Figure 3.5: (a) PSNR and (b) SSIM for different number of frequency bands considered in zig-zag scan order for smooth luma 8x8 blocks in WCE images

The coefficients of the DTT are quantized and entropy coded based on the block mode. A smooth block transformed using DTT results in insignificant high frequency coefficients which are reduced to zero after quantization. Therefore, only low frequency components in zig-zag order are considered for quantizing and computational cost of quantizing high components can be reduced by prior block type detection. Greater PSNR and SSIM are achieved when more coefficients are considered as low. But after a certain number of low frequency coefficients, the quality becomes constant and there is no significant improvement. Around 500 images from the esophagus to the colon are analysed. Based on this, around 12 coefficients for a luma 8x8 smooth block and 6 coefficients for 4x4 chroma block in zig-zag scan order are quantized,



(a) (b)

Figure 3.6: (a) PSNR and (b) SSIM for different number of frequency bands considered in zig-zag scan order for chroma 4x4 blocks in WCE images

while remaining are set to zero. It has been observed that considering more coefficients, doesnot exhibit significant improvement in quality as shown in Figure 3.5 and Figure 3.6. JPEG quantization table for $QF = 4$ is approximated for multiplication

4	2	2	4	8	8	16	16
4	4	4	4	8	16	16	16
4	4	4	8	8	16	16	16
4	4	4	8	16	16	16	16
4	4	8	16	16	32	32	16
8	8	16	16	16	32	32	16
16	16	16	16	32	32	32	32
16	16	16	32	32	32	32	32

(a) Luma

4	2	8	16
4	4	8	16
4	8	16	32
16	16	32	32

(b) Chroma

Figure 3.7: Modified JPEG Quantization at $QF = 4$ requires only bit-shifts

and addition free quantization. The quantization table shown in Figure 3.7 uses only bit-shift operations and reduces the computational cost incurred by multiplications and additions.

3.3.4 Coefficient Encoding

Coefficients obtained from quantization are entropy encoded in zigzag order using a low-complexity and low-memory encoder suitable for WCE. Zigzag ordering maps 8x8 transformed and quantized block into 1x64 one dimensional sequence starting from low frequency coefficients to high frequency coefficients. This type of reordering contributes to the increase in compression performance of the run-length entropy

coding schemes (Wallace (1992)).

The encoder algorithm uses run-length encoding. In this approach, non-zero AC coefficients are adaptive Golomb Rice (AGR) encoded using (3.3) and run of zero coefficients are Exponential Golomb coded (EGC) using (3.5). DC coefficients of neighbouring blocks exhibit strong correlation. Therefore, the difference between the adjacent DC coefficients is entropy coded using an AGR encoder.

$$GR(x,k) = [q \text{ zeros, } 1, \ k \text{ least significant bits of } x] \tag{3.3}$$

where

$$q = \left\lfloor \frac{x}{2^k} \right\rfloor \tag{3.4}$$

$$EGC(x) = [kzeros, 1, binary(x,k)] \tag{3.5}$$

where k is computed by

$$k = \lfloor \log_2(x+1) \rfloor \tag{3.6}$$

Since AGR can encode only non-negative integers, mapping function $M(res_{dc})$ given in (3.7) is used to convert difference residue to non-negative integers.

$$M(res_{dc}) = \begin{cases} 2res_{dc}, & res_{dc} \geq 0 \\ 2|res_{dc}|-1, & res_{dc}<0 \end{cases} \tag{3.7}$$

where,

$$res_{dc} = current_{DC} - previous_{DC} \tag{3.8}$$

where, $current_{DC}$ is the DC coefficient of the current block considered for encoding and $previous_{DC}$ is the DC coefficient of the previously encoded block.

The algorithmic description for encoding the coefficients is shown in Algorithm 3.1. In the algorithm Q_b is the block of quantized coefficients and B_m is the block texture mode. If $B_m == 0$, the block is considered as smooth otherwise non-smooth. if $C_r == 0$ the block is considered as luma otherwise chroma block. The AC coefficients of a smooth block are encoded in two steps. The cluster of first twelve low frequency coefficients are scanned in zigzag order and converted into a sequence of pair of elements (z,v), where v is the non-zero AC coefficient and z is the number of run of zero-valued coefficients preceding v. The signed elements in v are mapped to

Algorithm 3.1: Algorithmic description of quantized coefficients encoding

$function\ bitstream = entropy\ encode(Q_b, B_m, C_r)$
$b_s = ""$
$res_{dc} = current_{dc} - prev_{dc}$
$b_s = GR(M(res_{dc}), k_{dc})$
$[k_{dc}, N_{dc}, A_{dc}] = update(k_{dc}, res_{dc}, N_{dc}, A_{dc}$
if $C_r == 1$ **then**
  $N == 6$
else
  if $(B_m == 0)$ **then**
    $N == 12$
  else
    $N == 64$
  end if
end if
$z = 0$
for $n = 2 : N$ **do**
  $v = Q_b(n)$
  if $v == 0$ **then**
    $z \leftarrow z + 1$
  else
    $b_s = [b_s, EGC(z)]$
    $b_s = [b_s, GR(M(v), k_v)]$
    $[k_v, N_v, A_v] = update(k_v, abs(v), N_v, A_v)$
    $z = 0$
  end if
end for
if $v == 0$ **then**
  $b_s = [b_s,' EoB']$
end if
$bitstream = b_s$
end

non-negative integers using (3.9).

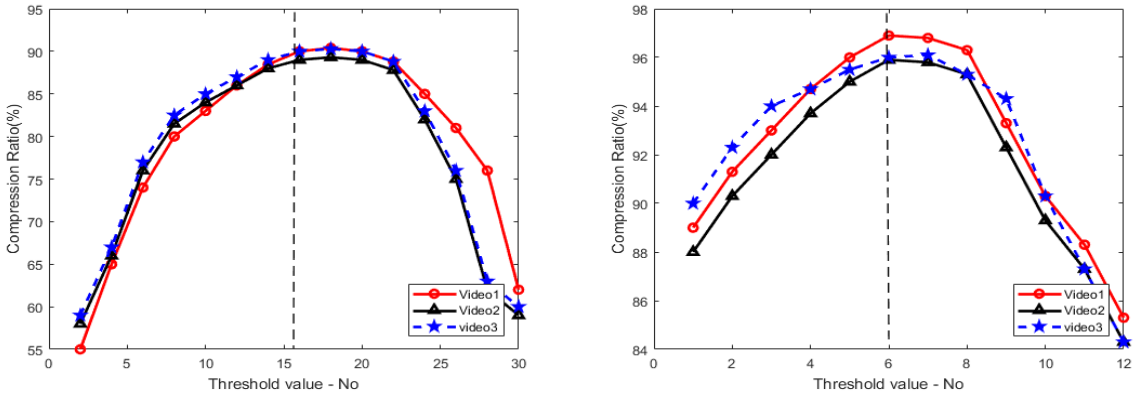
$$M(v) = \begin{cases} 2v - 1, & v>0 \\ 2(|v| - 1), & v<0 \end{cases} \tag{3.9}$$

All of the remaining majority symbols are almost zero in the smooth block. Therefore after scanning the first few coefficients, to indicate the end of the block, a single symbol EOB='111' is used. By using pre-identification of the block type, computations required for scanning and encoding of all the insignificant coefficients are reduced which saves computational power. Quantized coefficients in non-smooth blocks are encoded in the same fashion as smooth blocks except all the elements in a block are

scanned in zigzag order. Magnitude of z coefficients is always non-negative and small, therefore mapping is avoided for converting. Instead of AGR encoding ((Rice, 1979)) of z coefficients, EGC is used which is ideal for encoding small values. Initial value of k is computed using (3.6). Initial N_c and A_c values are set to zero for both res_{dc} and v coefficients. Similar procedure is followed for encoding chroma component. Parameter k is estimated using the method described in (Memon (1998)).

Algorithm 3.2: k-parameter updating based on values N_c and A_c.

$function\ [k, N_c, A_c] = update(k, coeff, N_c, A_c)$
$d = abs(coeff)$
while $N_c 2^k \geq A_c$ **do**
  $k \leftarrow k + 1$
end while
$N_c \leftarrow N_c + 1$
$A_c \leftarrow A_c + d$
if $N_c > N_0$ **then**
  $N_c \leftarrow \frac{N_c}{2}$
  $A_c \leftarrow \frac{A_c}{2}$
end if
end



(a) (b)

Figure 3.8: Compression efficiency of (a) Luma component and (b) Chroma component as function of N_o

Method for updating k based on N_c and A_c is as given in Algorithm 3.2. Threshold value N_o=16 for luma and N_o=6 for chroma constitutes a good estimation accuracy of k as shown in Figure 3.8a and Figure 3.8b. It has been observed that compression efficiency is maximum around the considered threshold values.

3.4 WZ Frame Encoder

3.4.1 WZ Coding of Luma Component

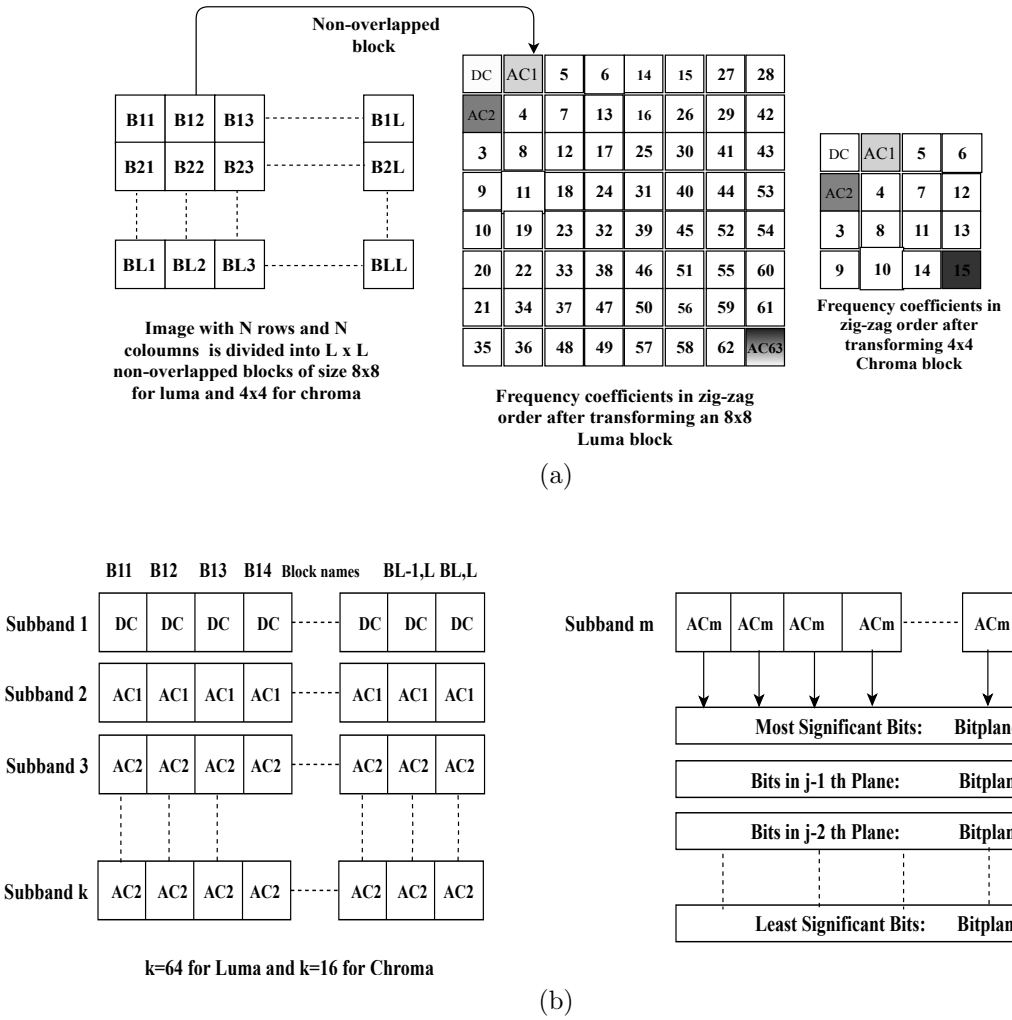


Figure 3.9: (a) Subband formation from frequency components of each block and (b) Bitplane extraction from subbands

Luma component obtained from colour space conversion is transformed into the frequency domain by applying 8x8 DTT on non-overlapped blocks. In WCE images, most of the region is smooth with uniform pixel distribution and applying smaller transform, e.g. 4x4 results in poor RD performance. Another advantage of using larger transforms is reduction in complexity of bit-plane encoding. For smaller transform, the frame results in more blocks with less number of subbands of larger bitplane length. Higher bitplane length need more number of gates to generate parity bits, resulting in higher complexity. Parity bit computation complexity can be minimized

by reducing the bitplane length which is possible by increasing the block size. From the quantized subbands, lower three subbands (DC, AC1 and AC2) are intra coded using AGR encoding which is described in Section 3.3.4.

Process of subband formation and bitplane extraction are shown in Figure 3.9a and Figure 3.9b. Bitplanes are extracted based on the maximum absolute value within each frequency band from remaining high frequency bands. From each subband j bitplanes are extracted where $j = \log_2[max(suband)]$. High quality SI is generated, when more low bands are intra coded. However, with an increase in the number of low bands, only fewer high bands are left to take advantage of WZ coding thus decreasing the performance. Therefore, only three components DC, AC1 and AC2 are intra coded and the remaining are considered as high-frequency bands which are WZ encoded.

Each bitplane in the subband is Slepian-Wolf encoded. Slepian-Wolf coding provides lossless encoding of two correlated frames. It achieves optimal bit-rate by independent encoding and joint decoding of two correlated frames. It consists of LDPCA encoder and a buffer to store parity bits. Extracted bitplanes from quantized coefficients are encoded independently to generate parity bits using LDPCA encoder which are stored in parity bit buffer. The generated parity bits serve as error correcting information at the decoder. These parity bits are transmitted in chunks upon reception of the request signal from the decoder. The low bands which are transmitted to the decoder assists in generating better quality SI, as the temporal correlation of low frequency bands is high.

3.4.2 WZ Coding of Chroma Component

Non-overlapped blocks of subsampled chroma component in 4:2:0 format is transformed into frequency domain using 4x4 DTT. Larger transform is not required for domain conversion as the chroma component has less homogeneous region due to subsampling. The chroma subband length is same as luma subband of length N (for k=1 to 16), which facilitates using same LDPCA encoder to generate parity bits. This concept avoids using another LDPCA encoder of different rate. All the chroma subbands are Slepian-Wolf encoded and transmitted in a similar fashion of luma component without splitting into low and high bands.

3.4.3 Side Information Generation and Decoding

The intra coded low bands of the luma component of the WZ frame is received first at the decoder side. Previously decoded frame is used for SI generation for WZ encoded frequency bands. Received DC, AC1 and AC2 frequency components from each block are entropy decoded. Motion estimation for each block is performed by using a block matching algorithm (BMA) considering the previously decoded frame. The BMA divides the current frame into macroblocks and compares each of the macroblock with the co-located block and its neighbouring blocks in the previously reconstructed frame. The three low frequency components from each block of WZ frame is matched with transformed blocks of the similar size in the previous frame. Since only reconstructed DC, AC1 and AC2 coefficients of luma component of WZ frame are available at the decoder side, the best match is selected based on minimum mean squared error (MSE). The MSE of two blocks X and Y with three low bands is calculated as given in (3.10).

$$MSE = (X_{DC} - Y_{DC})^2 + (X_{AC1} - Y_{AC1})^2 + (X_{AC2} - Y_{AC2})^2. \tag{3.10}$$

where, X corresponds to the transformed block of WZ frame and Y is the transformed block in the previously reconstructed frame located in the search area.

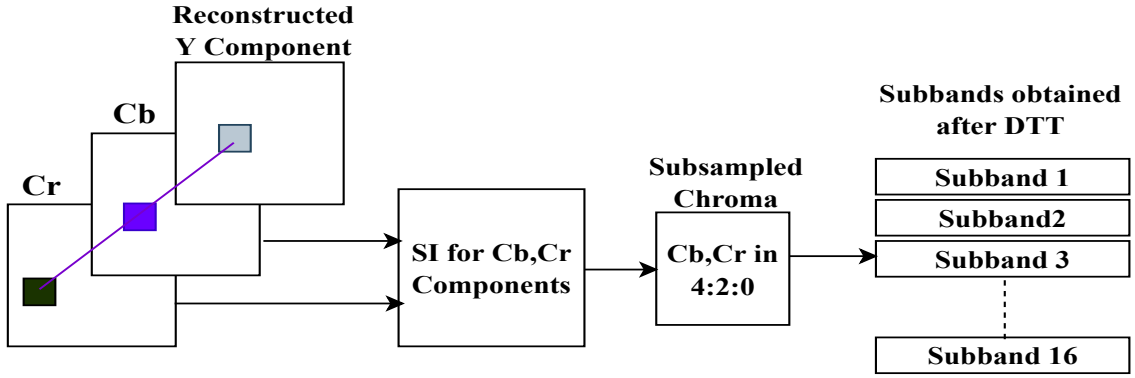


Figure 3.10: Side Information generation for chroma components from Intra-coded luma lowbands

The motion compensated frame is the new SI for the remaining WZ coded frequency bands. Motion estimation with Full search algorithm in a ±7 pixel search area is used in SI generation. The co-located block in chroma component of the previous frame is considered as best match for SI creation in case of WZ chroma as shown in Figure 3.10. Luma lowbands are used to search the best matching block with in the

search area. After the best match is found, its co-located chroma pixels can be used for the creation of chroma SI. Once the SI is found for chroma, Cb and Cr components are down-sampled and transformed using DTT to find the chroma SI frequency bands. SI is used to decode the remaining luma bands and chroma bands. SI refined using Slepian-Wolf decoder is reconstructed and stored in the frame buffer which is used for SI reconstruction of the future WZ frames.

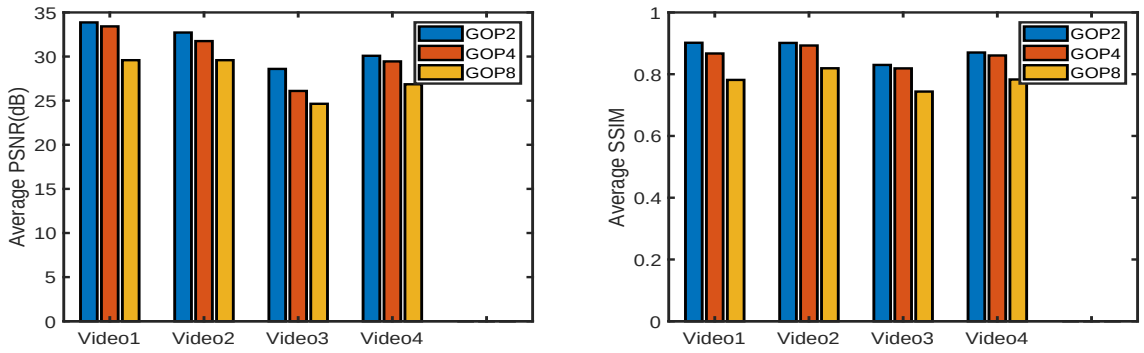
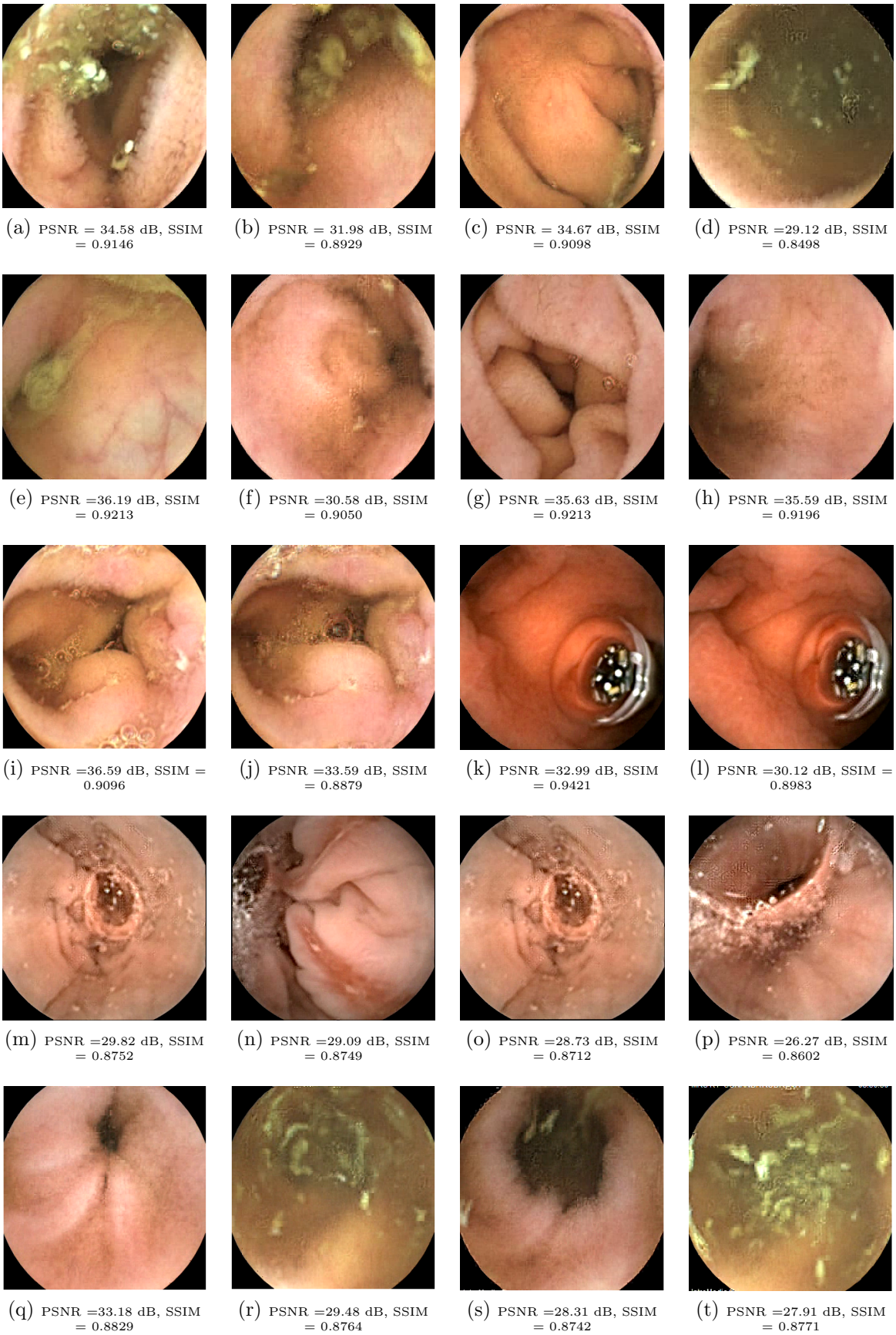


Figure 3.11: Average PSNR (dB) and SSIM of side information for different video sequences at GOP=2 ,4, 8

Quality of the SI is a very important parameter in deciding the performance of DVC techniques. To analyse the quality of SI generated, the four test videos of different motion characteristics are considered. Average PSNR and SSIM of SI for different WZ frames at different GOP size are shown in Figure 3.11. Quality of decoded WZ frames is better at GOP 2 compared to GOP=4, 8, but degrades the performance of the system due to increase in number of keyframes. In case of GOP 8, decrease in SI quality of higher WZ frames reduces the compression performance. GOP of 4 gives better SI quality with average PSNR of more than 30 dB for all WZ frames with one key-frame for every 4 frames. This method is expected to generate better SI quality compared to motion compensated frame interpolation. Table 3.1 shows a comparison between SI quality, generated by the DVC-FBC method and interpolation algorithm proposed by (Artigas *et al.* (2007)).

The visual quality of the SI frames for selected images is shown in Figure 3.12. SI is better for the video sequences with smooth translation and low motion, however can be poor due to absurd motion between frames when there is irregular and fast motion



(a) PSNR = 34.58 dB, SSIM = 0.9146 (b) PSNR = 31.98 dB, SSIM = 0.8929 (c) PSNR = 34.67 dB, SSIM = 0.9098 (d) PSNR =29.12 dB, SSIM = 0.8498

(e) PSNR =36.19 dB, SSIM = 0.9213 (f) PSNR =30.58 dB, SSIM = 0.9050 (g) PSNR =35.63 dB, SSIM = 0.9213 (h) PSNR =35.59 dB, SSIM = 0.9196

(i) PSNR =36.59 dB, SSIM = 0.9096 (j) PSNR =33.59 dB, SSIM = 0.8879 (k) PSNR =32.99 dB, SSIM = 0.9421 (l) PSNR =30.12 dB, SSIM = 0.8983

(m) PSNR =29.82 dB, SSIM = 0.8752 (n) PSNR =29.09 dB, SSIM = 0.8749 (o) PSNR =28.73 dB, SSIM = 0.8712 (p) PSNR =26.27 dB, SSIM = 0.8602

(q) PSNR =33.18 dB, SSIM = 0.8829 (r) PSNR =29.48 dB, SSIM = 0.8764 (s) PSNR =28.31 dB, SSIM = 0.8742 (t) PSNR =27.91 dB, SSIM = 0.8771

Figure 3.12: Visual quality of side information with PSNR and SSIM for frames of different test video sequences

Table 3.1: Comparison of SI quality between DVC-FBC and Motion compensated frame interpolation (MCFI) method for GOP of 2

	DVC-FBC		MCFI (Artigas *et al.* (2007))	
	Avg PSNR	Avg SSIM	Avg PSNR	Avg SSIM
Test Video1	33.86 dB	0.9418	26.23 dB	0.7830
Test Video2	33.22 dB	0.9013	25.98 dB	0.7734
Test Video3	29.59 dB	0.8502	22.45 dB	0.6931
Test Video4	31.08 dB	0.8702	23.23 dB	0.7112

of the capsule. Images with better SI require a less rate and bad SI need more rate. The SI generated by the DVC-FBC method is close to 35 dB which is the minimum threshold for acceptable medical image quality as suggested by the study (Istepanian *et al.* (2008)) which enables transmission of fewer parity bits for decoding and in turn saves transmission energy. SI which is an estimate of the WZ frame along with transmitted parity bits is used to reconstruct the WZ frame. Reconstructed WZ frame is stored in frame buffer after inverse quantization and inverse transform.

3.5 Complexity Analysis

3.5.1 Complexity Analysis of Keyframe Encoding

Table 3.2: Computations required per 8x8 block for transformation and quantization

Functional Block	Conventional Method	Proposed Method
Image Transformation	Integer DCT	Approximate DTT
	512 Additions, 224 Bit-shift Operations	384 Additions, 96 Bit-shift Operations
Quantization	JPEG Qunatization	Modified JPEG Quantization
	64 Multiplications	64 Bit shift operations

The complexity of the keyframe encoder and JPEG baseline encoder are evaluated with respect to the average number of computations needed for processing an 8x8 block. The analysis of complexity reduction is summarized in Table 3.2 and Ta-

Table 3.3: Computation reduction for an 8x8 block with block mode decision for BT-KFE

Functional stage	Without block mode decision	With block mode decision
Quantization	64 Bit-shifts	12 Bit-shifts for smooth, 64 Bit-shifts for non- smooth
Coefficient Encoding	64 comparisons for generating pairs of non-zero and zero coefficients	12 comparisons for coefficient encoding in case of smooth block
Reduction in computations with the texture based keyframe encoder		(0.81) x (% of number of smooth blocks)

ble 3.3. Using the block texture conditioned keyframe encoder, significant reduction in computations is achieved compared to JPEG baseline encoder. JPEG standard uses Huffman tables to encode the pairs of non-zero AC coefficients with a run of preceding zero coefficients. Huffman encoding requires two passes through the image which is very complex for WCE application and need extra memory for storing Huffman tables which can be the bottleneck for hardware implementations. The coefficients are fed to the encoder in the first pass. Huffman tables are created based on the frequency of occurrence of quantized coefficients. Huffman coding generates different tables for DC and AC coefficients of luma and chroma components (Wallace (1992), Pennebaker and Mitchell (1992)). In the second pass encoding of data using tables as reference is done. Constructing Huffman tables need extra computations which increases the complexity and also demand more memory. Along with encoded data, even tables need to be transmitted to the decoder which increases the power required for transmission. Proposed algorithm uses adaptive Golomb Rice encoding of quantized coefficients which does not employ table creation.

3.5.2 Complexity Analysis of WZ Frame Encoding

To develop a low complexity encoder, the following techniques are incorporated in the proposed DVC architecture.

- Elimination of hash creation and transmission: Capsule endoscopy video exhibits irregular motion. Therefore, sending hash to the decoder as extra infor-

mation helps in better SI generation which improves the rate-distortion performance. Hash based DVC architectures (Aaron *et al.* (2004), Deligiannis *et al.* (2012*a*, 2011), Ascenso *et al.* (2010)) result in better performance compared to conventional DVC architectures. But creating hash on the encoder side at block or frame level requires complex computations and it is extra overhead for the transmission. The DVC-FBC system does not employ any method to create and code the hash data. Instead, only a few low frequency bands of luma component are considered as intra bands that are used as hash to estimate SI at the decoder. The intra coded low bands are again not WZ coded and there is no extra overhead in the transmission of data. Thus, the proposed method reduce the encoding complexity compared to other hash based TDWZ-DVC architectures.

- Complexity reduction in WZ frame computational elements: Encoding of WZ frame in DVC-FBC comprises integer approximation of DTT, quantization, extraction of bit-planes and parity bit generation by LDPCA encoding. These computational elements need low computations compared to conventional TDWZ-DVC computational elements. DTT offers reduction in computations compared to DCT used in conventional TDWZ-DVC architectures. Quantizer need only bit shift operations. LDPCA is performed by just EX-ORing of a binary array with parity-check matrix bits. Very low amount of memory is required to store parity matrix bits since this matrix is sparse. Bit-plane length of chroma components is same as luma component. Therefore, separate LDPCA encoder of different length is not required to encode chroma bit-planes.

- Reduced latency in SI creation at the decoder: For an application such as WCE, the encoder complexity needs to be low, but the complexity of the decoder can be high. In the DVC-FBC system, a decoder with the feedback channel is implemented to control the rate. Generated parity bits are accumulated in a parity buffer are transmitted in chunks whenever the decoder sends the request signal. Parity bits received by the decoder are used for SI refinement. When the SI refinement is not satisfied, decoder requests encoder to transmit another chunk of parity bits. Parity bits are retained in the buffer till the decoder stops sending request signals to the encoder. This introduces delay in decoding the

WZ frame and limits the frame transmission rate. When another WZ frame is available for encoding, encoder cannot send the available WZ frame for decoding before completing the decoding of previous frame. So, encoder is compelled to store more parity bits in the buffer which increases the buffering complexity. Total delay in decoding of WZ frame depends on many factors (Deligiannis *et al.* (2012*b*)) and is given in (3.11).

$$T_d = \frac{t_{fa} + t_{SI} + F \times t_{tr} + F \times t_{LDPC}}{t_{fa}} \tag{3.11}$$

where t_{fa} = frame acquisition period, t_{SI} = time to generate side information, t_{tr} = time for transmission, t_{LDPC} = LDPC decoding time and F = number of feedback requests from the decoder.

In the proposed system, latency at the decoder is low since the encoding of the bitplanes to generate parity bits and SI creation on the decoder side happen simultaneously. Parity bits transmission in chunks of 25 bits to the decoder starts immediately after all the subbands are encoded. t_{SI} can be ignored as the encoding and transmission of intra bands of the WZ frame happen for each block whenever the transform of the block is available. Also, SI is generated with the simple motion compensation process on the previously decoded key-frame and it does not take much time. The time for transmission t_{tr} can be ignored as it is in the order of 15ns (Andreuccetti (2012)). Therefore, latency in decoding is due to the number of feedback requests F and t_{LDPC}. The calculation of frame rate (f_r) depends only on F and t_{LDPC} as given in (3.12). The total delay in decoding the frame is much reduced compared to (Boudechiche *et al.* (2017)) and (Deligiannis *et al.* (2012*a*)) as the time to generate SI is negligible.

$$f_r = \frac{1}{T_d} = \frac{1}{F \text{ x } t_{LDPC}} \tag{3.12}$$

The time required to LDPC decode a codeword of length of maximum 1944 bits with 27 to 81 iterations is $t_{LDPC} = 6\mu s$ per iteration (Brack *et al.* (2007)). For a frame resolution of 320 x 320 with 8 x 8 transform, the bit-plane length is 1600 bits. To transmit 1600 bits with the chunk length of 25 bits, 64 chunks are needed. Calculation of total number of chunks and time to decode the bitplanes of luma and chroma are given in Table 3.4. In luma, DC band, first

and second AC bands are intra coded. There are maximum 215 bit-planes of remaining subbands for luma, 48 for C_b and 48 for C_r components (calculated based on content based quantization). In the worst case, when entire bit-plane is transmitted at once, around $2ms$ is required to decode a bit-plane. Transmitting in 64 chunks, results in F=19904 for luma and chroma. To decode all the chunks, the decoder needs just 0.12s time in worst case with a frame rate of 8 fps. The frame rate can be greater than 8 fps because the decoder terminates its operation before requesting all the chunks, once all the bits are corrected. Also, frame rate can be further increased by decoding the bitplanes in parallel when a high frame rate is required.

Table 3.4: Computation of time required to decode the luma and chroma bitplanes of length=1600 in worst-case scenario

Component	#WZ bands	# bit-planes	n=Number of chunks per component	Time required to decode bitplanes in each component=$n \times t_{LDPC}$
Y	61	215	13760	82.56 ms
C_b	6	48	3072	18.3 ms
C_r	6	48	3072	18.3 ms
Total	73	311	19904	119.42ms
Total Number of chunks in one WZ frame F=19904			Time required to decode all the chunks=119.42 ms	

Encoder of the WZ frame computes all the parity bits and stores in fixed size buffer. These bits are transmitted to the decoder whenever encoder receives the feedback request from the decoder. Buffer size need not be changed if the decoder can complete decoding according to the frame capture rate. If the decoder can complete its operation before another frame is available at the encoder (frame decoding is faster than frame rate), decoder can wait until another frame is available. If the decoder cannot complete decoding within the available time, then an additional buffer at the encoder is required to store the

syndrome bits of the next WZ frame. However, because the required frame rate for the WCE application is just 10 fps, this method suffices (more than 10 fps) without employing an additional buffer due to the availability of SI before WZ decoding. Therefore, the DVC-FBC system can work at the higher frame rate and high resolution with decoding by using sufficient resources mainly parallel decoding of bit planes.

3.6 Simulation Results

The DVC-FBC system is evaluated for four test video sequences captured at different areas of the GI tract by Mirocam-Intromedic capsule which exhibits different motion characteristics. The performance evaluation is done by plotting average frame quality versus the average bitrate. The frame quality is measured in PSNR and SSIM. Medical image compression applications demand minimum reconstruction quality of at least 35 dB. Therefore, DVC-FBC is evaluated for GOP size of 2 and 4.

3.6.1 Performance Evaluation of BT-KFE

Table 3.5: Quality (PSNR) in dB and compression rate of the WCE luma images for the key frame encoder. (Q_{jpeg} is JPEG quantization table, $Q_{Modified}$ is modified quantization table)

Image	Without block mode decision			Smooth blocks in %	With block mode decision		
	PSNR (dB)		CR(%)		PSNR (dB)		CR(%)
	Q_{jpeg}	$Q_{Modified}$			Q_{jpeg}	$Q_{Modified}$	
Img 1	40.33	40.29	89	82	40.25	40.21	90
Img 2	42.79	42.76	87	72	42.77	42.74	89
Img 3	40.14	40.08	84	42	39.98	39.89	84
Img 4	40.30	40.28	84	45	40.22	40.19	84

To assess the performance of the keyframe encoder with block classification, images with smooth and non-smooth textural properties are considered. Img1 and Img2 has more number of smooth blocks and only low frequency coefficients in luma and

chroma blocks are encoded which saves bits required to encode the few high frequency coefficients. Ignoring higher coefficients from encoding slightly reduces the quality with improvement in CR. Img3 and Img4 consists less number of smooth blocks and all the frequency coefficients of most of the blocks are considered for encoding. Therefore, there will not be any improvement in CR. The compression-quality performance of the keyframe encoder without and with block classification is given in Table 3.5.

Table 3.6: Performance comparison between BT-KFE and other compression methods for WCE

Method	PSNR	SSIM	CR(%)
DPCM (Malathkar and Soni (2019))	∞	1	60.28
DPCM (Fante *et al.* (2016))	34.85	0.9913	86.34
DPCM (Khan and Wahid (2011*a*))	∞	1	67.47
DPCM+Subsampling (Khan and Wahid (2011*b*))	37.27	0.9964	77.14
DPCM (Chen *et al.* (2009))	42.24	0.9976	62.95
DCT (Turcza and Duplaga (2013))	35.21	0.9842	83.05
DCT (Lin and Dung (2011*a*))	25.31	0.9581	93.99
DCT (Wahid *et al.* (2008))	28.19	0.9606	89.06
DCT (Lin *et al.* (2006))	31.77	0.9646	89.46
DTT (BT-KFE)	37.14	0.9915	89.73

By applying block classification, complexity will be reduced without forgoing the quality of the reconstructed signal. As seen from the table, CR can be improved by 1-2% without change in PSNR for images with higher % of smooth blocks. The block classification technique is more useful in compression of endoscopic images due to smooth and homogeneous structure of GI tract. Table 3.6 compares the compression performance of the BT-KFE with the other compression methods. From the table, it can be observed that the BT-KFE provides better quality for the achieved CR compared to the other methods with lower complexity. This method outperforms the DPCM methods in terms of CR for nearly same quality and the DCT based methods in terms of quality for nearly the same CR.

3.6.2 RD Performance of DVC-FBC

The assessment of the DVC-FBC system is done by comparing the experimental results for YCbCr sequences with Motion JPEG, TDWZ-DVC and H.264/AVC intra (Main profile) at 8 fps. H.264 is very important for comparison as it is recognized codec for medical video compression (Yu *et al.* (2005)). The original test video sequences are in RGB format converted to YCbCr 4:2:0 format for testing using H.264 intra and TDWZ based DVC codecs.

The graphical comparison of RD results of the DVC-FBC system with benchmark codecs is shown in Figure 3.13. It can be observed that the performance of the DVC-FBC is better than Motion JPEG, but still less efficient compared to H.264-Intra. Though, H.264-Intra performs better than the DVC-FBC, intra prediction used in H.264-Intra is complex and not suitable for an application such as WCE. The results are also compared with MCFI based TDWZ codec. The DVC-FBC with frequency band classification gives better performance. RD performance of the DVC-FBC system with GOP=4 is better than GOP=2.

3.6.3 Bjontegaard-Delta Metrics

The Bjontegaard delta (BD) rate savings and PSNR improvement introduced by the DVC-FBC compared to MJPEG, TDWZ-DVC and H.264-Intra for GOP=4 and GOP=2 is presented in Table 3.7 and Table 3.8. DVC-FBC performs better than MJPEG in terms of BD rate savings of around 60% with PSNR gain of 8 dB for GOP=4 and 42% with PSNR gain of 5 dB for GOP=2. Compared to TDWZ-DVC, DVC-FBC achieves BD rate savings of 40% and 20% with PSNR gain of 5 dB and 2 dB for GOP=4 and 2 respectively. The gain in compression performance is achieved at reduced encoder complexity. H.264-Intra performs better than DVC-FBC and gives better BD rate savings of around 45% with PSNR improvement of 3 dB.

3.6.4 Encoding Time

The encoding time depends on two components: WZ coding and the keyframe coding. The encoder complexity is determined by the encoding time for the entire sequence in seconds. The encoding time can give a reasonably accurate estimation of the encoder

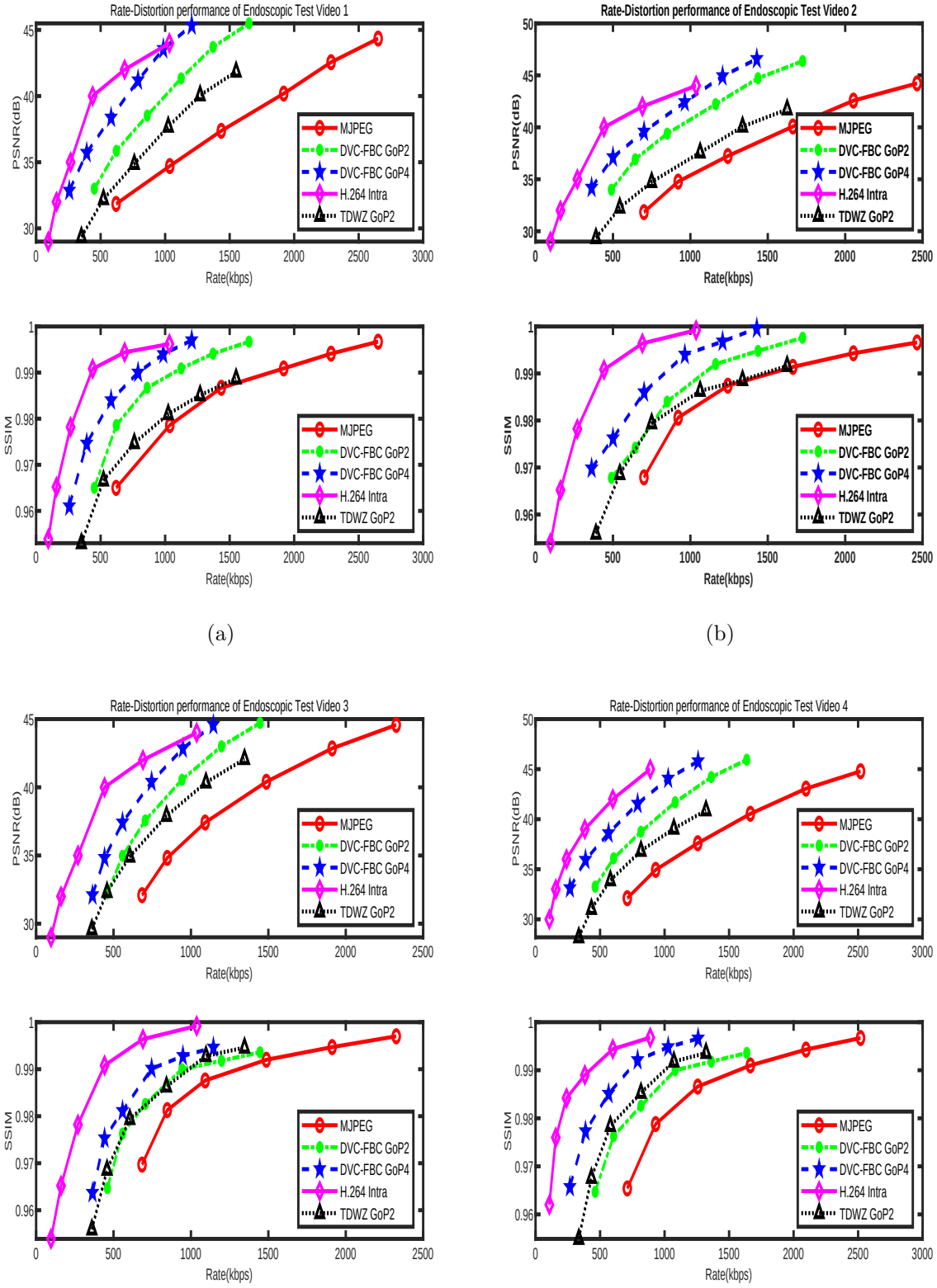


Figure 3.13: Rate-distortion performance for 320 x 320 endoscopic test video sequences with 8 frames/second

Table 3.7: The BD bit-rate saving, PSNR gain in dB and SSIM gain of the DVC-FBC with GOP=4 compared to other coders

Test Video sequences	MJPEG			TDWZ-DVC			H.264-Intra		
	BD (%) bit-rate	BD PSNR	BD SSIM	BD (%) bit-rate	BD PSNR	BD SSIM	BD (%) bit-rate	BD PSNR	BD SSIM
Video1	-63.86	+8.48	+0.0173	-49.72	+5.75	+0.0150	+39.42	-2.39	-0.0102
Video2	-56.20	+7.59	+0.0122	-47.77	+5.60	+0.0098	+47.87	-2.73	-0.0124
Video3	-49.18	+6.93	+0.0103	-28.68	+3.34	+0.0053	+60.69	-3.12	-0.0114
Video4	-59.72	+8.22	+0.0136	-46.32	+5.32	+0.0064	+51.39	-3.01	-0.0122

Table 3.8: The BD bit-rate saving, PSNR gain in dB and SSIM gain of the DVC-FBC with GOP=2 compared to other coders

Test Video sequences	MJPEG			TDWZ-DVC			H.264-Intra		
	BD (%) bit-rate	BD PSNR	BD SSIM	BD (%) bit-rate	BD PSNR	BD SSIM	BD (%) bit-rate	BD PSNR	BD SSIM
Video1	-46.54	+5.62	+0.0110	-24.32	+2.52	+0.0077	+77.14	-5.50	-0.0146
Video2	-44.25	+5.66	+0.0059	-32.46	+3.56	+0.0026	+88.55	-4.59	-0.0187
Video3	-40.89	+5.22	+0.0061	-11.97	+1.27	+0.0060	+82.42	-4.37	-0.0138
Video4	-42.05	+5.45	+0.0068	-19.57	+2.07	+0.0030	+79.77	-6.02	-0.0174

complexity under appropriate simulation conditions. Table 3.9 compares the encoding time of the DVC-FBC and time reduction computed using (3.13) over the reference encoders at four different RD points.

$$\text{Time Reduction in } \% = \frac{T_{Ref} - T_{DVC-FBC}}{T_{Ref}} X100 \tag{3.13}$$

Where, T_{Ref} is the encoding time of the reference encoder in seconds and $T_{DVC-FBC}$ is the encoding time of the DVC-FBC in seconds.

The RD points selected are at different quantization parameters which encodes the video sequences at approximate bitrates close to 300, 600, 1000 and 1500 kbps. Time reduction indicates by how much the complexity of the proposed codec is reduced compared to reference codec. Positive reduction indicates that the proposed encoder requires less time and negative reduction indicates the proposed consumes more time compared to reference encoders. The average encoding time per frame computed

Table 3.9: Comparison of encoding time and time reduction by the proposed (DVC-FBC) over the reference encoders at various bitrates

Test sequences	Bitrate (kbps)	Encoding Time in seconds				Time reduction in% over reference encoders		
		MJPEG	TDWZ-DVC	H.264-Intra	DVC-FBC	MJPEG	TDWZ	H.264-Intra
Video-1	500	57	132	240	62	-8.77	+53.03	+74.16
	1000	58	132	241	66	-15.72	+50.21	+72.61
	1500	58	168	306	78	-36.84	+53.65	+74.50
	2000	59	190	346	80	-40.25	+57.95	+76.87
Video-2	500	52	115	210	56	-7.69	+51.52	+73.33
	1000	53	127	232	60	-13.20	+52.98	+74.13
	1500	53	146	267	66	-24.52	+55.06	+75.28
	2000	53	156	303	71	-33.35	+57.40	+76.56
Video-3	500	41	92	168	45	-9.75	+51.30	+73.21
	1000	42	102	186	47	-11.90	+54.06	+74.73
	1500	42	117	214	54	-28.57	+54.12	+74.76
	2000	43	132	241	54	-25.58	+59.26	+77.59
Video-4	500	44	99	180	48	-9.01	+51.52	+73.33
	1000	45	110	200	51	-15.90	+53.64	+74.50
	1500	45	154	280	57	-29.54	+62.99	+79.64
	2000	45	176	320	61	-38.63	+65.34	+80.93

+ and - indicates complexity reduction and increase in proposed compared to reference encoders respectively

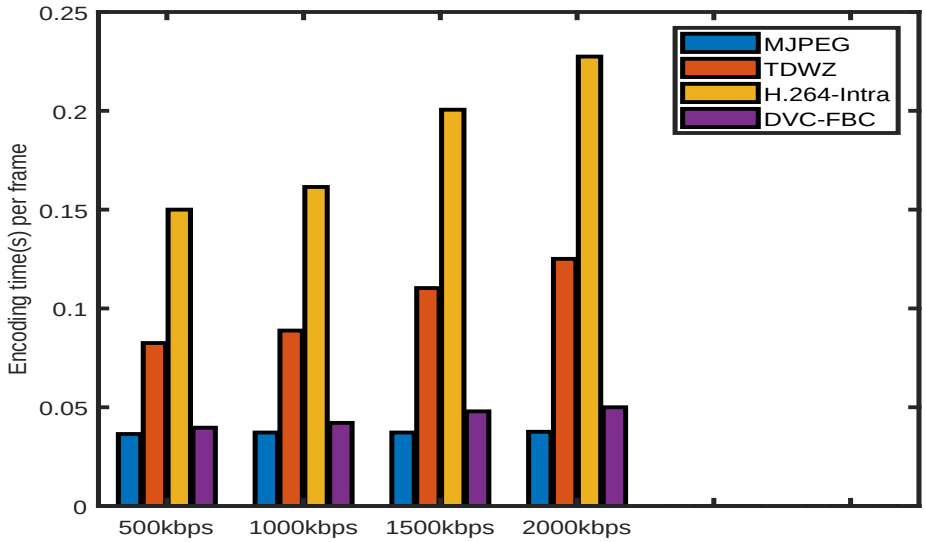


Figure 3.14: Comparison of DVC-FBC encoder complexity with reference encoders, averaged over all the frames in test video sequences

over all the four video sequences at different bit-rates is provided in Figure 3.14. Proposed DVC-FBC encoder needs an average encoding time of 0.0420 seconds per

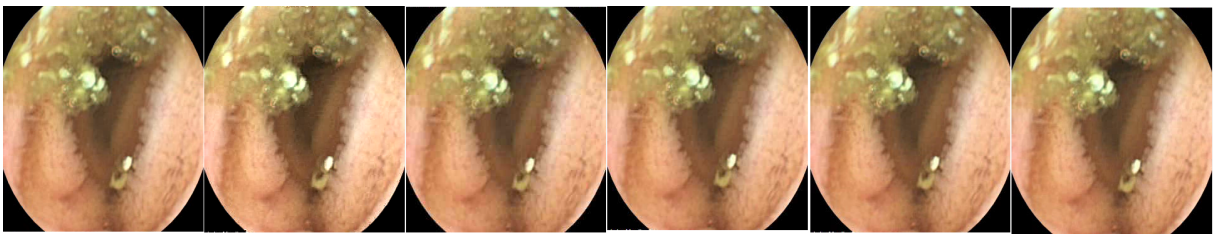
frame. Though MJPEG needs 0.0372 seconds per frame which is faster than DVC-FBC, it results in poor RD performance. H.264-Intra consumes average encoding time of 0.2 seconds per frame and substantially much complex than DVC-FBC due to RD optimization. TDWZ need more time as 50% of the frames are keyframes and H.264-Intra coded.

3.6.5 Visual Performance

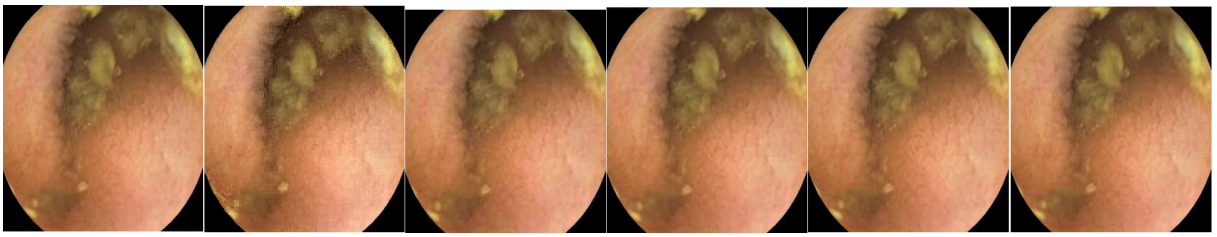
The visual quality of the reconstructed WZ frames of the DVC-FBC codec for GOP=4 with different rates along with original and SI frames is shown in Figure 3.15. SI frame has very good quality in some regions and distorted in some regions. Distortion can be corrected just after decoding the first few subbands and thus require less bitrate. But finer details of the texture are clearly visible only after decoding all the subbands. It can be observed that higher quality requires higher rate. Medical imaging demands greater than 35 dB image quality which requires more than 800kbps rate when the frame transmission rate is 8 frames per second.

3.7 Summary

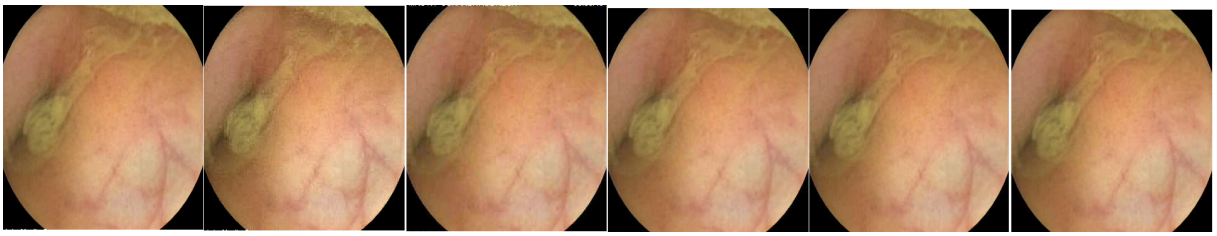
Motivated by low complexity encoder requirements for WCE video application, a simple encoder based on distributed video coding that exploits temporal correlation with novel texture conditioned key-frame encoder is designed. Most of the WCE video content has insignificant high frequency coefficients due to homogeneous and smooth textured regions of the GI tract. Texture conditioned key-frame encoder reduces the complexity of quantizing and entropy coding of insignificant frequency components using a simple block classifier for the smooth region without compromising on the quality. To combat the irregular motion characteristics of WCE video content, the DVC-FBC system intra codes lower frequency coefficients of each 8x8 luma block which acts as a hash to generate good quality SI at the decoder. This reduces the complexity of hash creation and transmission. The generated SI quality is high and can go upto 36 dB for the video sequences with low motion characteristics. A novel approach for WZ coding of subsampled chroma component and generation of SI for chroma is introduced. Experimental results show that, this system outperforms



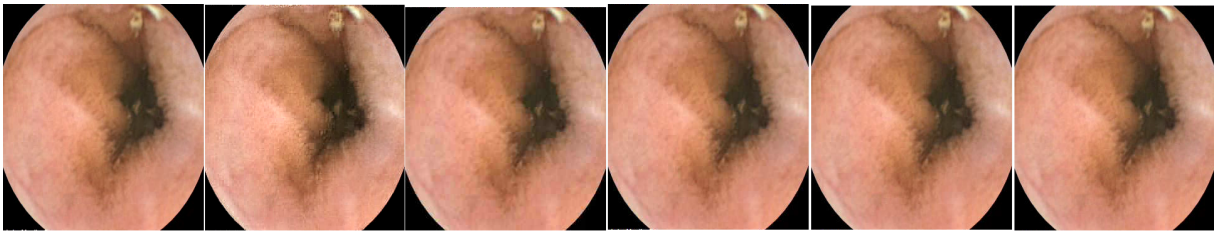
(a) Endoscopic Test Video1; (i) Original, (ii) SI frame and reconstructed frames at (iii) 622kbps; 33.71 dB; 0.9786 (iv) 861kbps; 36.37; 0.9867 (v) 1124kbps; 40.18 dB; 0.9909 (vi) 1371kbps; 41.55 dB; 0.9941



(b) Endoscopic Test Video2; (i) Original, (ii) SI frame and reconstructed frames at (iii) 644kbps; 33.77 dB; 0.9806 (iv) 850kbps; 36.23; 0.9874 (v) 1163kbps; 39.07 dB; 0.9914 (vi) 1437kbps; 41.56 dB; 0.9943



(c) Endoscopic Test Video3; (i) Original, (ii) SI frame and reconstructed frames at (iii) 508kbps; 33.82 dB; 0.9813 (iv) 655kbps; 36.91; 0.9876 (v) 892kbps; 40.18 dB; 0.9920 (vi) 1147kbps; 42.85 dB; 0.9947



(d) Endoscopic Test Video4; (i) Original, (ii) SI frame and reconstructed frames at (iii) 605kbps; 34.13 dB; 0.9788 (iv) 817kbps; 36.59; 0.9866 (v) 1082kbps; 39.54 dB; 0.9910 (vi) 1362kbps; 42.05 dB; 0.9943

Figure 3.15: Visual performance of 320 x 320 endoscopic test video sequences at 8 frames/second with PSNR and SSIM index of SI frames decoded at different rates for GOP=4

MJPEG by 60% and TDWZ-DVC by 40% in compression with an average encoding time of $42ms$ per frame. The DVC-FBC achieves better RD performance compared to TDWZ-DVC with 50% reduction in complexity.

Chapter 4

DVC Architecture with Deep Chroma Prediction Model

4.1 Introduction

The consecutive frames in the WCE video exhibits homogeneity in colour and texture. Considering this, it is sufficient to transmit only the luma components of the WZ frames while the chroma components can be predicted using keyframes. In WCE, decoding is done using a powerful computing system with no memory and power constraints. This chapter presents a DVC architecture with a CNN based deep neural network to predict chroma components of the WZ frame from keyframe at the decoder. This eliminates the processing of WZ-chroma components thus reducing the computational complexity with improved compression performance. The proposed deep neural network for chroma prediction achieves better prediction performance with less computation time. The deep chroma prediction model consists of a merging block with a spatial attention mechanism to merge the feature maps extracted from reference and target frames. So, the model can exploit the non-local similarities between the two frames. Thus, the chroma prediction model can efficiently transfer color from the keyframe to similar regions in the WZ frame.

4.2 Proposed DVC Architecture with Deep CNN (DVC-DCP) at the Decoder

A short sequence of frames captured in a particular GI organ has a lot of similarity in colour and texture. Due to similarity in colour and textural characteristics it is sufficient to transmit only the luma components of WZ frames and chroma components

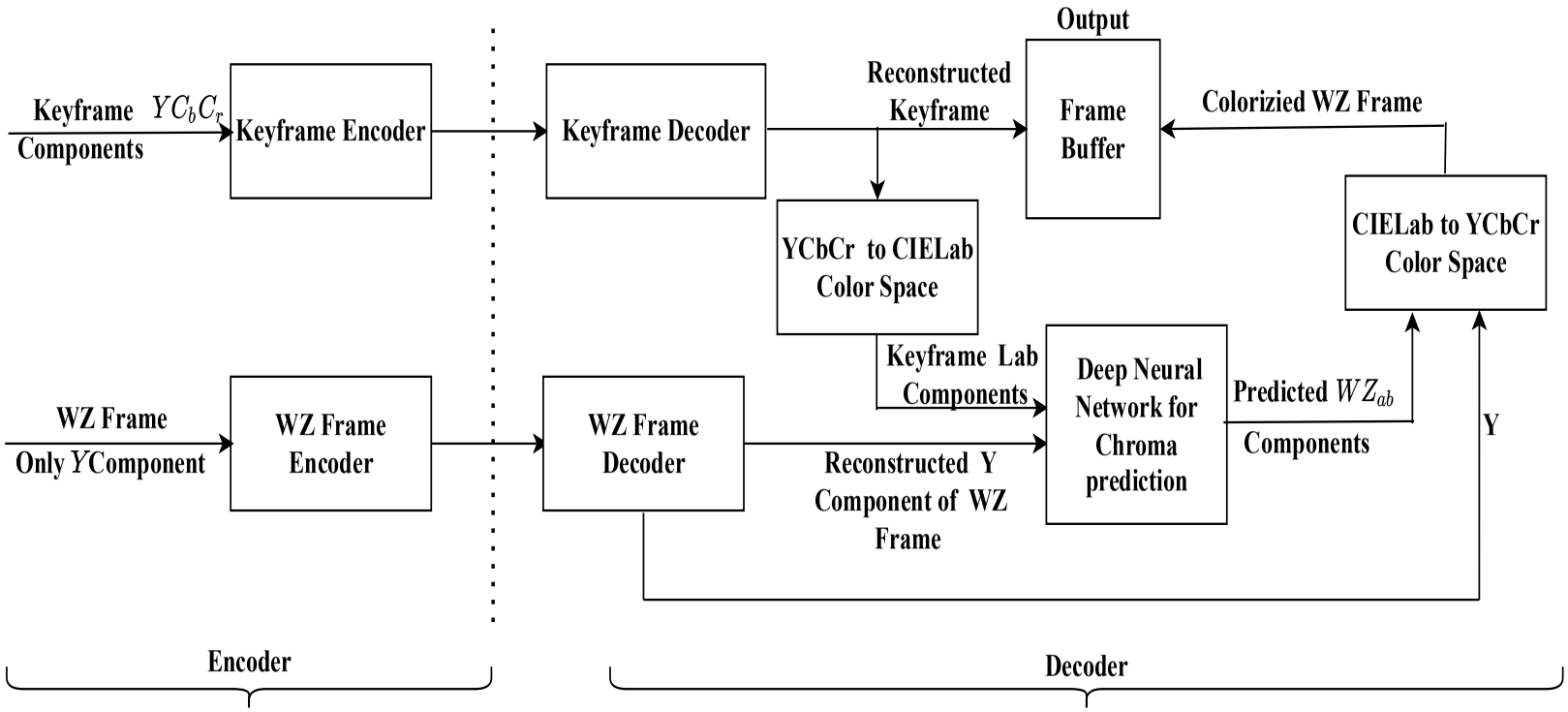


Figure 4.1: Proposed DVC architecture with CNN for WZ-chroma generation on the decoder side

can be generated using keyframes. The modified DVC framework which consists of a deep CNN at the decoder side to predict chroma components of WZ frame is shown in Figure 4.1. Y, C_b and C_r components of keyframe and Y component of WZ frames are encoded using the technique proposed in Chapter 3. At the decoder, the keyframes are decoded and stored in the frame buffer. Only Y component of the reconstructed keyframes is used for SI generation of Y component of the WZ frame. The reconstructed WZ frame Y component and the keyframe from the same GOP are given as inputs to the deep chroma prediction model. The proposed compression technique, intends to reduce the encoder complexity by encoding only the luma components of WZ frames.

4.2.1 Deep Chroma Prediction Model

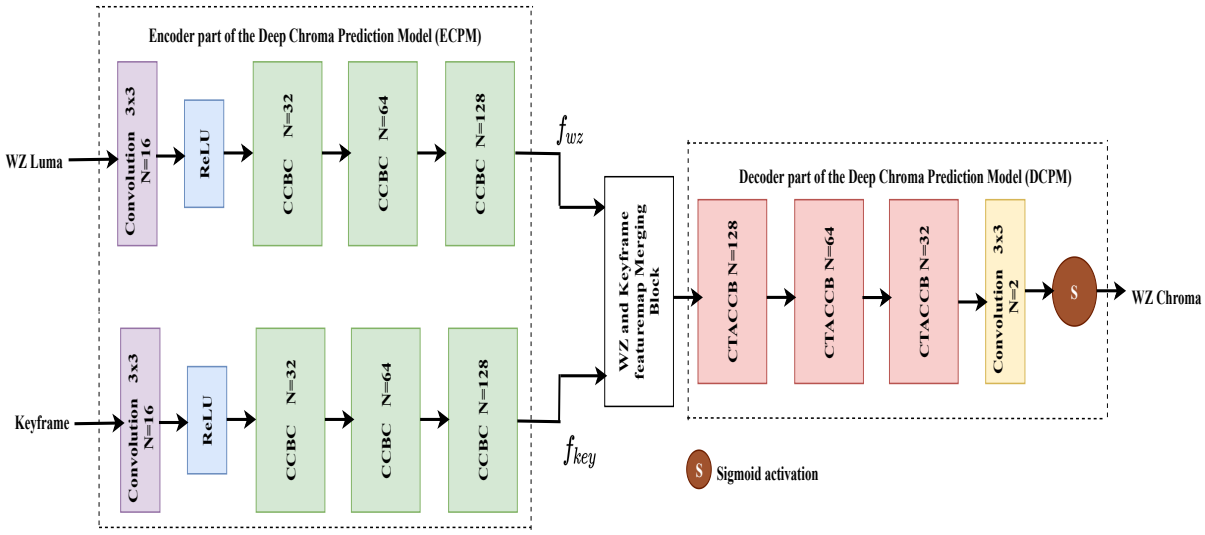


Figure 4.2: Proposed Deep CNN architecture for WZ frame chroma prediction

Many chroma prediction techniques are proposed for general images and videos, where the chroma of the reference frame is transferred to the gray scale target frame by mapping luma and related texture information of the small rectangular image patches within an image (Welsh *et al.* (2002), Yatziv and Sapiro (2006)). Some methods generate chrominance maps for the target image by using global colour statistics of the source such as histogram, mean and variance (Freedman and Kisilev (2010)). Since these methods ignore spatial pixel details, they yield improper results. Pixel, superpixel and segment levels of colour correlation are considered in other set of

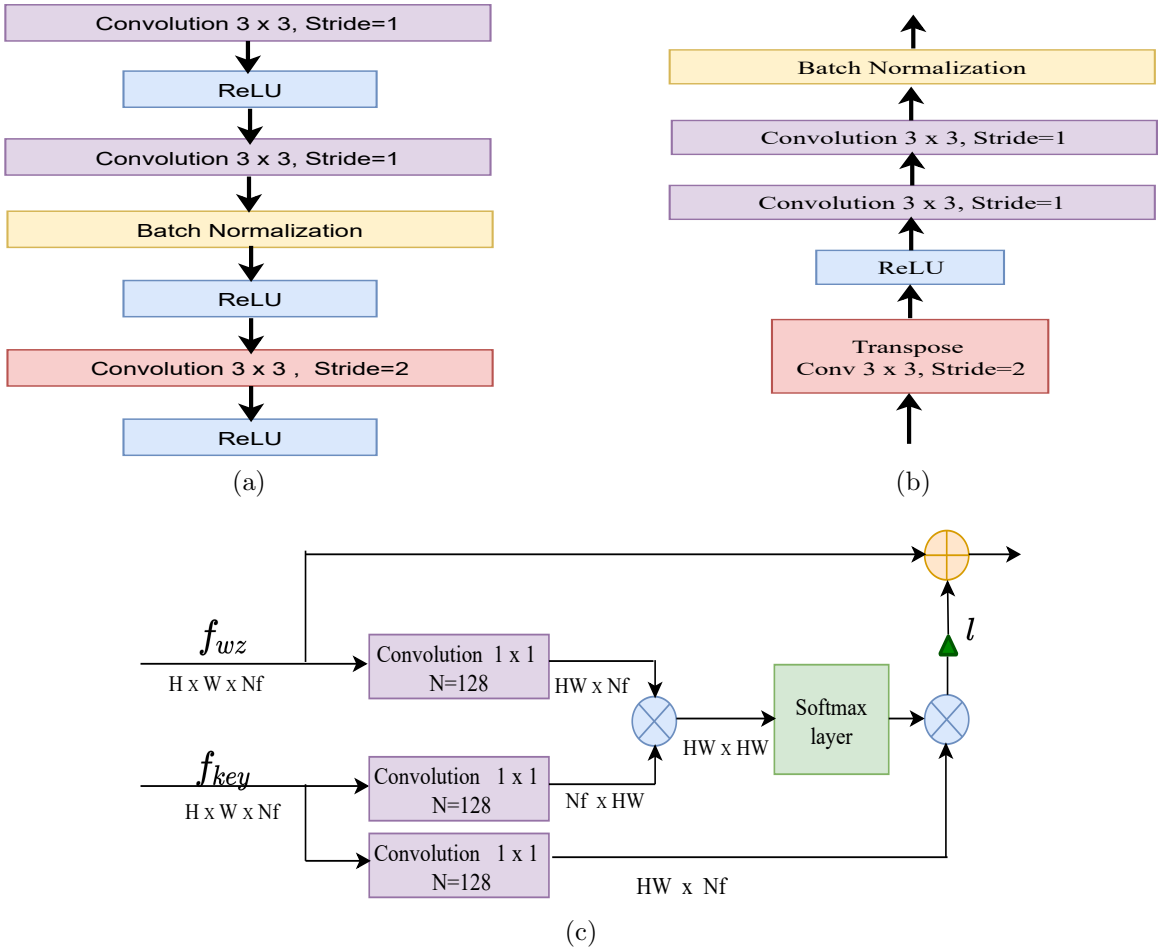


Figure 4.3: Main building units of the CNN architecture (a) CCBC unit (b) CTACCB (c) Merging block, H and W is the height and width of the feature maps, Nf is the number of feature maps

approaches (Liu *et al.* (2008)). Dictionary-based methods are proposed for chroma map generation using a reference image. The target image is colourized using a chroma map by matching luma pixels with the colour map dictionary (Khan *et al.* (2016)). The main problem in the above conventional colour mapping approaches lie in the selection of a reference patch in the source image for generating the correct colour map. It is very difficult to find an appropriate matching for all the image patches in the reference image. This leads to a lot of problems in medical diagnosis as the colour information is very important.

To overcome the issues in above methods, a deep-learning based CNN architecture shown in Figure 4.2 is proposed to generate appropriate chrominance map for the WZ frame by accepting a keyframe as reference and WZ frame as a target input image. Later the map is fused with the luma component of WZ frame to obtain the final

colourized WZ frame. In the figure, N represents number of kernels in each layer. f_{key} and f_{WZ} are the keyframe and WZ frame features. Detailed representation of the main components used in the architecture is shown in Figure 4.3

a. CIELab Colour Space: In literature, many studies have suggested that deep CNNs trained using images from a specific domain perform better when represented in a particular colour space. This is because, CNN learning and decision making depends on the input and converting an image from one colour space to another generates an entirely different input comprising different numbers. Motivated by this, the proposed CNN model is trained using images in YCbCr and CIELab colour space. Training the model using the other color spaces such as RGB and HSV is not possible, because at the decoder only Y component of the WZ frame is available. The model performs better in predicting the chroma when trained in CIELab colour space compared to YCbCr. Therefore, luma component of the WZ frame and keyframe luma and chroma components in CIELab colour space are given as inputs to the deep chroma prediction model. The prediction performance of the model trained using two different colour spaces measured in terms of PSNR and SSIM is given in Table 4.2 of Section 4.3.1. Images in CIELab colour space exhibits wider colour spectrum compared to other colour spaces and gives better performance. It also generates a better perceptually linear colour space and is ideal for training computer vision models due to its perceptual uniformity (Ishikura *et al.* (2017)). The L component is same as a Y component of YCbCr colour space. Hence, it is sufficient to predict only the a and b chroma components of CIELab space. All these features of CIELab colour space makes it more effective for training the deep learning models compared to other colour spaces Gowda and Yuan (2018).

b. Model Details: The deep chroma prediction model is based on deep CNN encoder and decoder based architecture, consisting of ten feature blocks. The model is trained to transfer the chroma of a similar region from keyframes to WZ frames. Deep CNN encoder and decoder based architecture is selected because this architecture is proven to be more effective in many image generation applications (Badrinarayanan *et al.* (2017)). The merging block connects the encoder part of the chroma prediction model (ECPM) and the decoder part of the chroma prediction model (DCPM). ECPM takes reconstructed keyframe components and reconstructed WZ frame luma

component as an input. DCPM extracts the features from the input, which are the compact representations of the input components. The merging block merges the relevant keyframe and WZ-frame spatial details using feature maps extracted by ECPM. This layer utilizes the non-localized texture and colour similarities between the reference keyframe and the WZ frame. Merged features are transmitted to DCPM. It will learn to generate the chrominance information by matching similar textural regions of the luma. This is performed by converting the combined key and WZ features to chrominance features in three stages using temporal convolutions. Thus the model learns to transfer the chroma to the similar areas of the WZ frame from the keyframe.

Table 4.1: Details of the deep colour-prediction model

(a) Reference (Keyframe) deep CNN encoder

Layer	# Filters	Filter Size	Output Size
Input	-	-	320 x 320 x 3
Conv-1	16	3x3	320 x 320 x 16
Conv-2	32	3x3	320 x 320 x 32
Conv-3	32	3x3	320 x 320 x 32
Conv-4	32	3x3	160 x 160 x 32
Conv-5	64	3x3	160 x 160 x 64
Conv-6	64	3x3	160 x 160 x 64
Conv-7	64	3x3	80 x 80 x 64
Conv-8	128	3x3	80 x80x1286
Conv-9	128	3x3	80 x80x128
Conv-10	128	3x3	40 x40x128

(b) Target (WZ) deep CNN encoder

Layer	# Filters	Filter Size	Output Size
Input	-	-	320 x 320 x 1
Conv-11	16	3x3	320 x 320 x 16
Conv-12	32	3x3	320 x 320 x 32
Conv-13	32	3x3	320 x 320 x 32
Conv-14	32	3x3	160 x 160 x 32
Conv-15	64	3x3	160 x 160 x 64
Conv-16	64	3x3	160 x 160 x 64
Conv-17	64	3x3	80 x 80 x 64
Conv-18	128	3x3	80 x80x1286
Conv-19	128	3x3	80 x80x128
Conv-20	128	3x3	40 x40x128

(c) Merging block

Layer	# Filters	Filter Size	Output Size
Input	-	-	40 x 40 x 128
Conv-21	128	1 x 1	40 x 40 x 128
Conv-22	128	1 x 1	40 x 40 x 128
Conv-23	128	1 x 1	40 x 40 x 128
$\oplus$	Matrix addition		
$\otimes$	Matrix multiplication		
l	l learnt constant parameter		

(d) Decoder part of chroma prediction model

Layer	# Filters	Filter Size	Output Size
Conv-Transpose-1	128	3x3	80 x 80 x 128
Conv24	128	3x3	80 x 80 x 128
Conv25	128	3x3	80 x 80 x 128
Conv-Transpose-2	64	3x3	160 x 160 x 64
Conv26	64	3x3	160 x 160 x 64
Conv27	64	3x3	160 x 160 x 64
Conv-Transpose-3	32	3x3	320x320x32
Conv28	32	3x3	320x320x32
Conv29	32	3x3	320x320x32
Conv30	2	3x3	320 x 320 x 3

The ECPM consists of a convolution layer and three CCBC blocks. C represents a convolution and B represents a batch normalization (BN) layer. CCBC block consists of two convolution layers followed by a BN layer. The output of the BN layer is reduced by half in size by a convolution layer with stride=2. The structure of the CCBC unit is shown in Figure 4.3a. DCPM consists of three CTACCB blocks. Where, CTA represents a transposed convolution layer, C represents convolution and B represents a batch normalization layer. Each CTACCB block consists of three convolution layers and a BN layer as shown in Figure 4.3b. The input to each block consists of up-sampled version of the features from the previous layer. The final layer of the DCPM consists of the convolution layer and sigmoid activation layer. These two layers convert the multi-level feature maps into two-level WZ frame chrominance map. The proposed deep colour prediction model consists of 33 convolution layers; 20 in the encoder path (conv1-conv20), 3 in the link connecting ECPM and DCPM (conv21-conv23) and 10 in the decoder path (conv24-conv33). The output of all the convolution layers is activated by the ReLU activation function. The details of convolution layers in terms of filter-size, the number of filters and the output size is described in Table 4.1.

c. Model Loss Function: To train the model both the luma component of the WZ frame (WZ_L), luma-chroma components of the keyframe (K_{Lab}) are fed to the CNN. The network predicts the chroma component of WZ frame with the trained parameters of the network θ as:

$$\widehat{WZ_{ab}} = CNN(K_{Lab}, WZ_L; \theta) \tag{4.1}$$

The predicted WZ chroma component $\widehat{WZ_{ab}}$ should be same as ground truth WZ_{ab} if the network transfers the chroma by proper texture matching. To evaluate the chrominance loss while training the model, the *smooth* L_1 distance is computed for every pixel and integrated over the entire image. As a distance metric, Smooth_L1 loss is considered that eliminates the averaging problem (Zhang *et al.* (2017)). The network consisting of parameters θ should learn to minimize the chrominance loss function given in (4.2).

$$Loss = \sum_{pixels} Smooth\_L_1\big(\widehat{WZ_{ab}}(pixel), WZ_{ab}(pixel)\big) \tag{4.2}$$

Where $Smooth\_L_1(x_1, x_2)$ is defined as :

$$Smooth\_L_1(x_1, x_2) = \begin{cases} \frac{1}{2}(x_1 - x_2)^2, & \text{if} \quad |x_1 - x_2| < 1 \\ |x_1 - x_2| - \frac{1}{2}, & \text{otherwise} \end{cases} \tag{4.3}$$

d. Dataset and Training Details of Chroma Prediction Model: Training dataset is generated by extracting around 10000 frame pairs from different WCE videos. Videos captured from different organs such as esophagus, stomach, small bowel, colon of GI tract are considered to generate dataset consisting keyframes, WZ frames and groundtruth. Frames extracted from these videos consists various GI anomalies such as intestinal bleeding, Crohn's disease, colon polyps etc., while others are normal without any abnormalities. Luma component of the WZ frames considered as target frames in this work are stored in WZ frame folder. Keyframe folder consists every fourth frame extracted from considered videos which is paired with all the WZ frames. Keyframes are considered as reference frames, from which colour is transferred to WZ frame by matching luma and texture by deep colour prediction model. Original WZ frames are treated as ground truth labels to compute chrominance loss.

The proposed CNN is trained with the batch size of 8 using adam optimizer. Model is trained in an entirely supervised style with a smooth L1 chrominance loss. Network parameters are initialized using He-normal initialization. Network implementation and training for 50 epochs is done using Keras, a deep learning API using Tensorflow as backend on an NVIDIA Tesla-T4 GPU. The initial learning rate is fixed to 0.0001 which decays for every 10 epochs by a factor of 2. Chrominance loss computed using (4.2) is used to update the network parameters during back propagation.

4.3 Simulation Results & Discussions

The proposed DVC-DCP is evaluated for four test video sequences given in Table 1.2. The performance of the compression system is evaluated by plotting the average bitrate versus average frame quality measured by PSNR in dB and SSIM. The colour similarity between the WZ frame after the colour transfer from the keyframe and the original WZ frame is computed by CIE76-ΔE colour difference and structure and hue similarity (SHSIM). CIE76-ΔE and SHSIM are computed using (4.4) and (4.5)

respectively.

$$\Delta E = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\sqrt{(L_o - L_c)^2 + (a_o - a_c)^2 + (b_o - b_c)^2} \tag{4.4}$$

where L, a, b are the CIE-Lab colour space components of the original and colourized image.

$$SHSIM(x_o, x_c) = \frac{SSIM(x_o, x_c) + 0.2H(x_o, x_c)}{1.2} \tag{4.5}$$

where $H(x_o, x_c)$ is the hue similarity computed between original image x_o and colourized image x_c using (4.6).

$$H(x, y) = \frac{2\lambda_x\lambda_y}{\lambda_x^2 + \lambda_y^2} \tag{4.6}$$

where the λ_x and λ_y is the mean hue of x and y.

4.3.1 Evaluation of Deep Colour Prediction Model

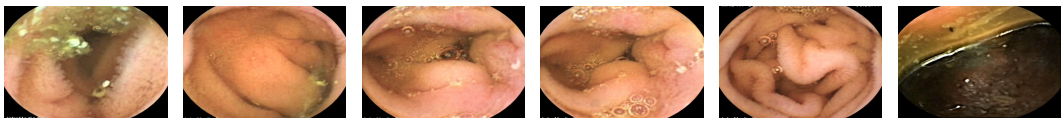
Table 4.2: Chroma prediction performance comparison for test video sequences in YCbCr and CIELab colour space

Video Sequence	Number of WZ frames	YCbCr				CIELab				T
		PSNR (dB)	SSIM	SHSIM	ΔE	PSNR (dB)	SSIM	SHSIM	ΔE	
Video-1	300	42.21	0.9894	0.9903	4.23	44.67	0.9956	0.9938	3.94	25
Video-2	263	40.34	0.9889	0.9911	4.86	42.23	0.9942	0.9926	4.50	25
Video-3	210	36.75	0.9923	0.9893	4.46	40.41	0.9962	0.9936	4.38	23
Video-4	225	37.27	0.9888	0.9864	3.49	38.98	0.9922	0.9892	3.15	23

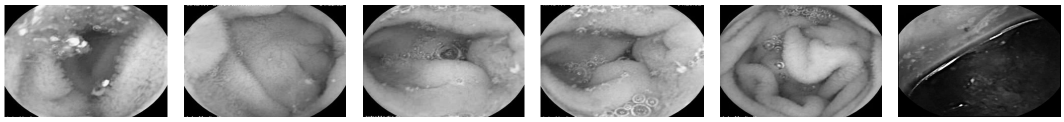
T: Time taken for predicting chroma components of the WZ frames in each video sequence in seconds

The proposed deep chroma prediction model is is tested on WCE videos captured at different locations of the GI tract. Colour transfer performance from keyframes to WZ frames for the test videos in YCbCr and CIELab colour space is measured by using PSNR, SSIM, SHSIM and ΔE. Average of the measurements along with processing time for the entire video sequence is given in Table 4.2. In comparison to YCbCr colour space, the chroma prediction model performs better in CIELab colour space.

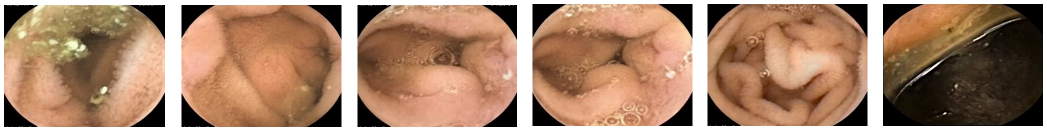
Visual performance of the chroma prediction model for pair of key and WZ frames which exhibits very less motion and high similarity is shown in Figure 4.4. It can



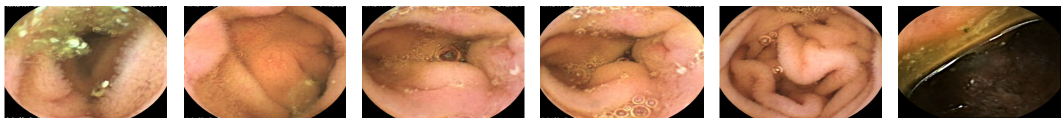
(a) Keyframes



(b) Gray WZ frame

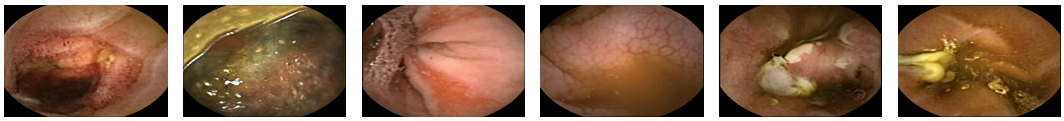


(c) WZ frames after colour transfer evaluated with PSNR, SSIM, SHSIM and ΔE

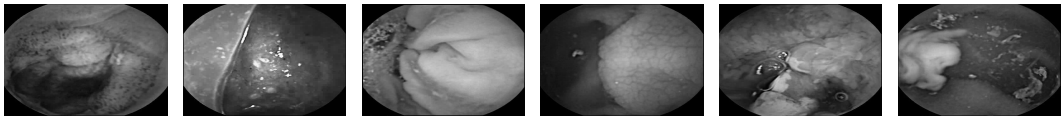


(d) Groundtruth WZ frames

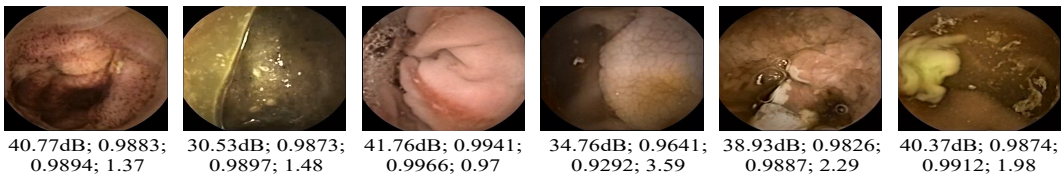
Figure 4.4: Visual performance of colour transfer between keyframe and WZ frame with very less motion



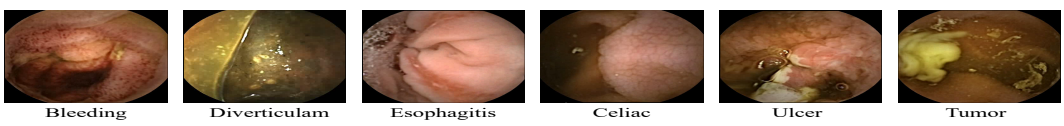
(a) Keyframes



(b) Gray WZ frame

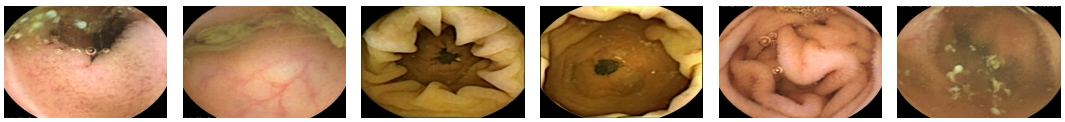


(c) WZ frames after colour transfer evaluated with PSNR, SSIM, SHSIM and ΔE

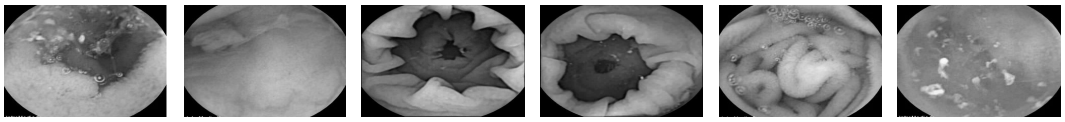


(d) Groundtruth WZ frames

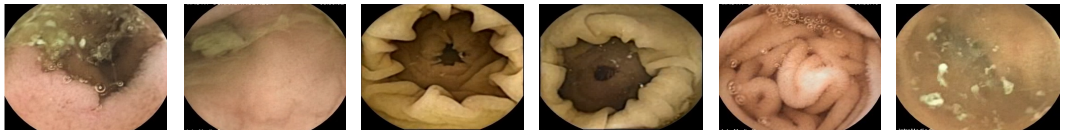
Figure 4.5: Visual performance of colour transfer between the frames with abnormalities



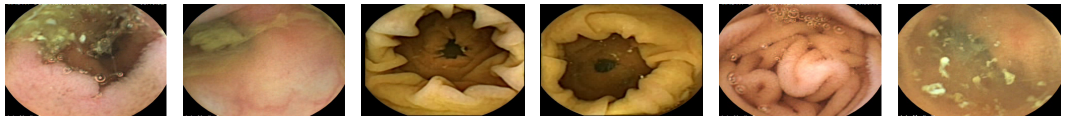
(a) Keyframes



(b) Gray WZ frame



(c) WZ frames after colour transfer evaluated with PSNR, SSIM, SHSIM and ΔE



(d) Groundtruth WZ frames

Figure 4.6: Visual performance of colour transfer between frames with fast-motion

be observed that colourized WZ frame is similar to original WZ frame and indicates accurate colour transfer by perfect matching of luma component and texture. Colour plays an important role in detection of some GI tract related abnormalities like bleeding, ulcer and inflammation (Li and Meng (2009), Yuan *et al.* (2015)). In Figure 4.5, visual performance of the colour transfer between frames with various anomalies is shown. It can be observed that bleeding, ulcerated and other lesioned image regions are perfect coloured. Colourized WZ frames shows high colour similarity with original WZ frames. When the capsule moves fast there will be a large variation between the consecutive frames. The visual performance of the colour transfer in such case is shown in Figure 4.6. A very slight degradation in the colour pertaining to these frames is observed that will not affect the diagnosis performance when image analysis is done by the physician.

4.3.2 Evaluation of DVC-DCP Architecture

The proposed DVC-DCP is evaluated by comparing the rate distortion (RD) performance and encoder complexity with MJPEG, TDWZ, DVC-FBC and H.264-Intra

codecs for the considered test video sequences. To assess the RD performance, PSNR and SSIM are plotted against the average bit rate. Encoder complexity is measured using encoding time. The encoding time can give a reasonably accurate estimation of the encoder complexity under appropriate simulation conditions.

A Encoding Complexity

Table 4.3: Comparison of encoding time and time reduction by DVC-DCP over the reference encoders at various bitrates

Test sequences	Bitrate (kbps)	Encoding Time in seconds					Time reduction in% over reference codecs			
		MJPEG	TDWZ	H.264-Intra	DVC-FBC	DVC-DCP	MJPEG	TDWZ	H.264-Intra	DVC-FBC
Video-1	500	57	132	240	62	47	+17.54	+64.39	+80.41	+24.19
	1000	58	132	241	66	50	+13.79	+62.12	+79.25	+24.24
	1500	58	168	306	78	66	-13.79	+60.71	+78.43	+15.38
	2000	59	190	346	80	68	-15.25	+64.21	+80.34	+15.00
Video-2	500	52	115	210	56	42	+19.23	+63.67	+80.00	+25.00
	1000	53	127	232	60	45	+15.23	+64.56	+80.60	+25.00
	1500	53	146	267	66	57	-7.54	+60.95	+78.65	+13.63
	2000	53	156	303	71	60	-13.20	+61.53	+80.19	+15.49
Video-3	500	41	92	168	45	35	+14.63	+61.95	+79.16	+22.22
	1000	42	102	186	47	35	+16.66	+65.68	+81.18	+25.53
	1500	42	117	214	54	48	-14.28	+58.97	+77.57	+11.11
	2000	43	132	241	54	48	-11.62	+63.63	+80.08	+11.11
Video-4	500	44	99	180	48	36	+18.18	+63.63	+80.00	+25.00
	1000	45	110	200	51	39	+13.33	+64.54	+80.50	+23.52
	1500	45	154	280	57	51	-13.33	+66.88	+81.78	+10.52
	2000	45	176	320	61	51	-13.33	+71.02	+84.06	+16.39

Positive reduction indicates the proposed encoder requires less time and negative reduction indicates the proposed consumes more time compared to reference encoder.

In DVC-DCP, the encoding of chroma components of WZ frames are eliminated, instead predicted at the decoder using keyframes as reference frames. In a video sequence encoded with GOP size=4, 75% of the frames are WZ frames, the chroma components of which are not encoded. This saves the time required for chroma processing and hence reduces the complexity of the encoder. Encoding complexity measured in terms of time taken to encode the entire sequence and complexity reduction of the proposed method as compared with MJPEG, TDWZ, H.264-Intra and DVC-FBC is given in Table 4.3.

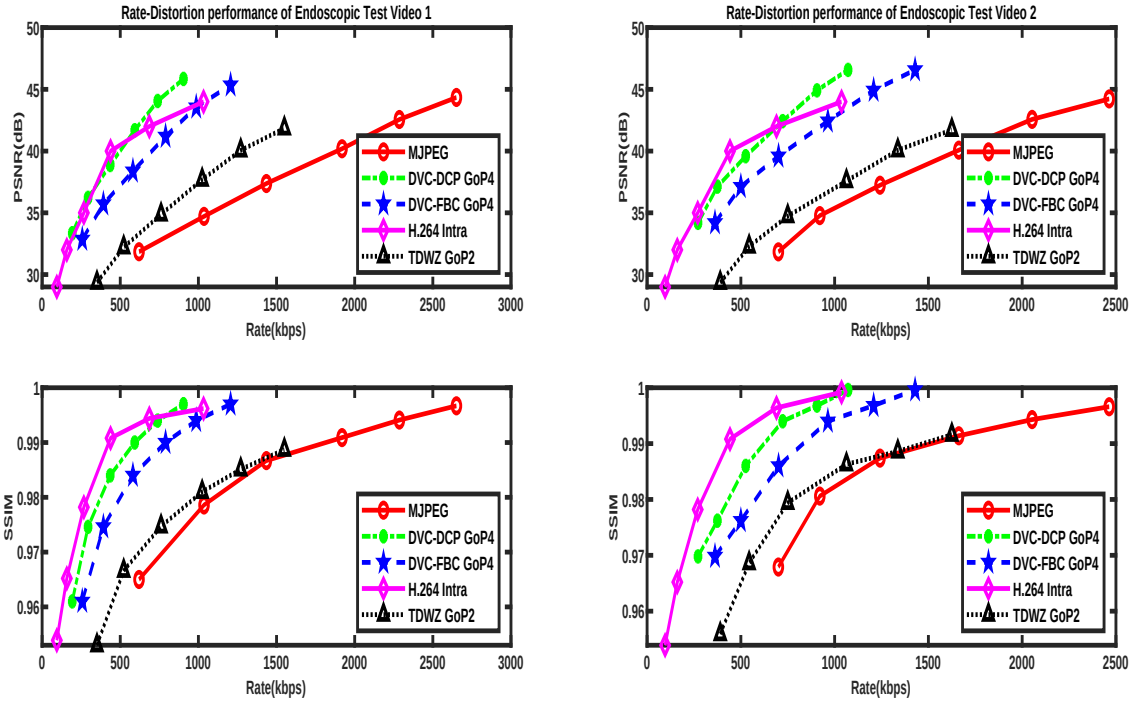
B Rate Distortion Performance

To assess the performance of the proposed DVC-DCP model, the graphical comparison of the RD results is done with MJPEG, TDWZ, H.264 Intra and DVC-FBC as shown in Figure 4.7. It can be observed that the performance of the proposed technique is much better than MJPEG, TDWZ and DVC-FBC and close to H.264 Intra.

Proposed method performs better in-terms of quality loss, compared to DVC-FBC which involves sub-sampled WZ frame chroma transmission along with luma. In DVC-FBC, WZ chroma subsampling is done before encoding in which only one chroma pixel is selected for every 4 pixels of WZ luma. At the decoder during the reconstruction of the transmitted image, the sub-sampled chroma component is up-sampled by interpolation of pixels. This is one of reason for the quality loss incurred after reconstruction. Another major cause for the quality loss is quantization at the encoder. Since the chroma component is not encoded in the proposed method, chroma reconstruction is not performed. Instead the chroma components are predicted at the decoder and the quality loss is introduced when the prediction is not accurate.

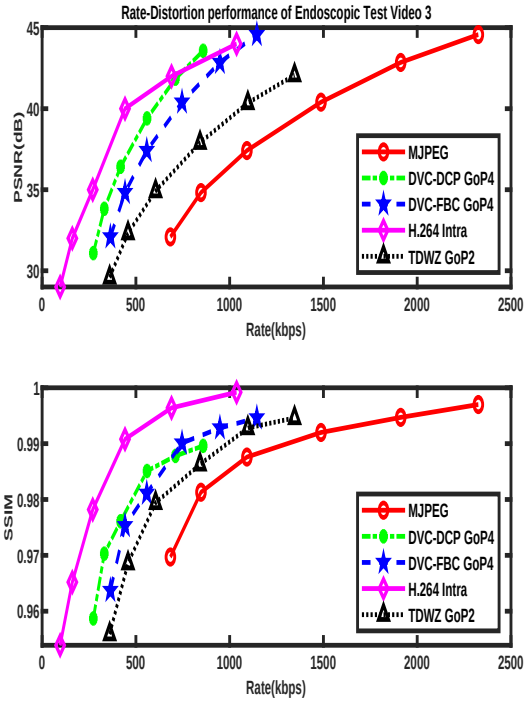
The quality comparison of chroma prediction by deep CNN model with chroma reconstruction using dequantization and upsampling for the test sequences is shown in Figure 4.8. From the comparison it can be observed that, deep chroma prediction method performs nearly same as chroma reconstruction. Slight quality degradation can be observed for the video sequences with fast motion. But in WCE, the 75% of the images are captured in stomach and small intestine where the capsule motion is very slow or sometimes shows no motion. Therefore the rate distortion performance of the proposed method with chroma prediction is better compared to DVC-FBC.

The DVC-DCP method achieves better Bjontegaard delta (BD) bit-rate savings than MJPEG and DVC-FBC. BD rate savings and improvement in PSNR over MJPEG and DVC-FBC is given in Table 4.4. BD rate savings and improvement in SSIM over MJPEG and DVC-FBC is given in Table 4.5. The savings in bit-rate are achieved at reduced encoder complexity. Compared to the proposed method, H.264 Intra performs better with PSNR improvement of 1.2 dB and bit-rate savings of around 20% for the video sequences captured in esophagus and colon. DVC-DCP performs same as H.264 Intra for the video captured in intestine which exhibits very

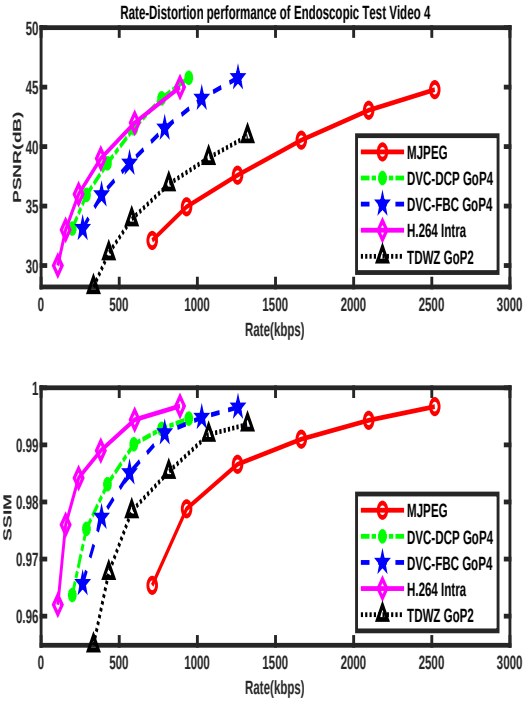


(a)

(b)



(c)



(d)

Figure 4.7: RD performance for test video sequences with 8 frames per second

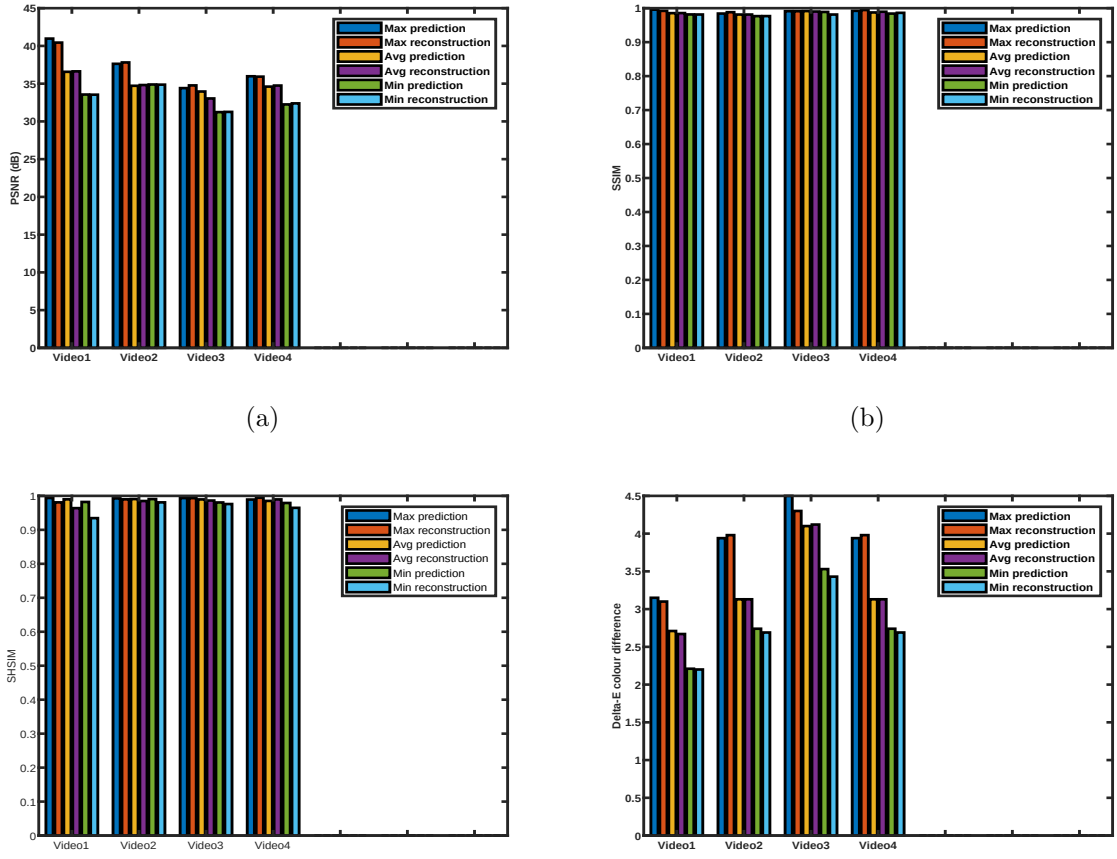


Figure 4.8: Quality comparison of WZ-chroma deep prediction and chroma reconstruction methods using (a) PSNR (dB), (b) SSIM (c) SHSIM and (d)ΔE. Lower ΔE indicates better performance

Table 4.4: The BD bit-rate savings in % and PSNR gain in dB of the DVC-DCP compared to MJPEG, DVC-FBC, TDWZ-DVC and H.264-Intra

Test Video	MJPEG		DVC-FBC		TDWZ-DVC		H.264-Intra	
Sequences	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR
Video1	-74.32	+11.22	-29.47	+2.83	-64.45	+8.63	+6.48	-0.21
Video2	-67.18	+10.32	-25.02	+2.52	-60.82	+8.05	+10.69	-0.45
Video3	-57.03	+8.95	-17.75	+2.06	-36.16	+4.25	+35.04	-2.06
Video4	-69.08	+10.95	-25.00	+2.34	-59.71	+7.74	+22.27	-1.09

Table 4.5: The BD bit-rate savings in % and SSIM improvement of the DVC-DCP compared to MJPEG, DVC-FBC, TDWZ-DVC and H.264-Intra

Test Video	MJPEG		DVC-FBC		TDWZ-DVC		H.264-Intra	
Sequences	Bitrate	SSIM	Bitrate	SSIM	Bitrate	SSIM	Bitrate	SSIM
Video1	-61.93	+0.0208	-15.36	+0.0037	-55.24	+0.0186	+19.08	-0.0071
Video2	-56.84	+0.0192	-25.09	+0.0067	-48.82	+0.0167	+13.51	-0.0065
Video3	-43.93	+0.0125	-11.68	+0.0023	-26.12	+0.0084	+39.55	-0.0117
Video4	-63.61	+0.0205	-17.38	+0.0057	-43.44	+0.0127	+17.95	-0.0081

slow changes from frame to frame. The RD results obtained are with reduced encoder complexity when compared to other encoding systems.

C Performance Comparison with WCE Image Compression Methods

Table 4.6: Performance comparison between proposed method at different bitrates and existing WCE image compression methods

Method	PSNR (dB)	SSIM	CR (%)
DPCM (Malathkar and Soni (2019))	∞	1	60.27
DPCM (Fante *et al.* (2016))	34.84	0.9912	86.33
DPCM (Khan and Wahid (2011*a*))	∞	1	67.46
DPCM+Subsampling (Khan and Wahid (2011*b*))	37.25	0.9962	77.14
DPCM (Chen *et al.* (2009))	42.23	0.9975	62.94
DCT (Turcza and Duplaga (2013))	35.22	0.9841	83.04
DCT (Lin and Dung (2011*b*))	25.32	0.9582	93.98
DCT (Wahid *et al.* (2008))	28.18	0.9605	89.05
H.264-Intra modified (Dung *et al.* (2008))	36.24	0.9828	82.12
DVC-FBC at 437kbps	33.18	0.9691	93.17
DVC-FBC at 575kbps	36.12	0.9757	91.02
DVC-FBC at 1039kbps	41.42	0.9931	83.77
DVC-DCP at 327kbps	34.18	0.9654	94.89
DVC-DCP at 582kbps	40.58	0.9885	90.91
DVC-DCP at 1128kbps	46.57	0.9988	82.37

The proposed work is also compared with other WCE image compression techniques in terms of PSNR and SSIM at different compression ratios. Individual frames in the video sequences are compressed using WCE image compression methods and

average CR, PSNR and SSIM is taken for comparison. To compare the proposed method with the image compression techniques CR is calculated using (4.7).

$$CR = 1 - \frac{E}{R} \tag{4.7}$$

Where,

E= Total bits transmitted from encoder to decoder,

R= $F \times M \times N \times$ Bits per pixel,

F= Number of frames in the video,

M, N= Number of rows and columns in each frame,

Average PSNR and SSIM is computed between each frames of the original video and reconstructed video for comparison of compression and quality is listed in Table 4.6. From the results it can be observed that the method provides better PSNR and SSIM at the high CR, maintaining low complexity encoder.

4.4 Summary

A low complexity DVC based WCE video compression algorithm with deep chroma prediction model at the decoder is proposed. The proposed chroma prediction model is trained to effectively transfer the colour from the keyframe to WZ frame using a spatial attention mechanism. It uses the inter-spatial correlation of feature maps extracted from the encoder part of the chroma prediction model to build a visual attention map and learns effectively to colourize similar areas of the target image using the chroma from the reference image. Also, the effect of colour space in training the deep CNN is illustrated. The proposed model is trained in $CIELab$ colour space and performs better than YC_bC_r colour space. Overall, the achieved quality of chroma prediction is relatively the same as chroma reconstruction.

In the DVC-DCP technique, the WZ-frame chroma components are ignored from encoding which improved the compression efficiency at reduced encoder complexity. In the DVC-DCP, SI generation is required only for luma. Since the entire chroma is predicted at the decoder, the SI quality does not have any impact on chroma reconstruction. This improved the quality with reduction in bit-rate. Chroma SI refinement is not done at the decoder which saves the decoding time and leads to an improvement in frame rate.

The proposed DVC-DCP method achieved better BD bit-rate savings than MJPEG, DVC-TDWZ and DVC-FBC. The video sequences are encoded with a GOP size=4 and the chroma components of the 75% of the WZ frames are not encoded which reduces the encoding complexity of the chroma components. The encoding complexity of the proposed DVC-DCP is almost same as MJPEG achieved at much lower bitrate. RD performance of the DVC-DCP model is better than DVC-FBC model by 25% in bitrate savings with PSNR gain 2.5 dB which is achieved at reduced encoder complexity. The DVC-DCP outperforms DVC-TDWZ in terms of RD and encoding complexity performance. The RD performance of the DVC-DCP is close to H.264 Intra achieved at much lower complexity.

Chapter 5

WCE Video Summarization

5.1 Introduction

This chapter proposes a WCE video summarization (WCE-VS) framework consisting of a convolutional autoencoder that can extract deep semantic features in an unsupervised way. The similarity measure in the extracted deep features is used to segment the video into shots. From each shot, the keyframes are extracted with the help of the motion profile obtained by inter-frame motion energy and direction. In WCE, the change in each frame is due to movement of the capsule. Therefore, it is possible to extract keyframes which covers the entire WCE video space of a shot with the help of motion analysis. The presented method achieves better summarization performance measured in terms of F-Score and compression ratio compared to the existing WCE video summarization techniques depend on handcrafted feature selection methods.

5.2 WCE Video Summarization Framework

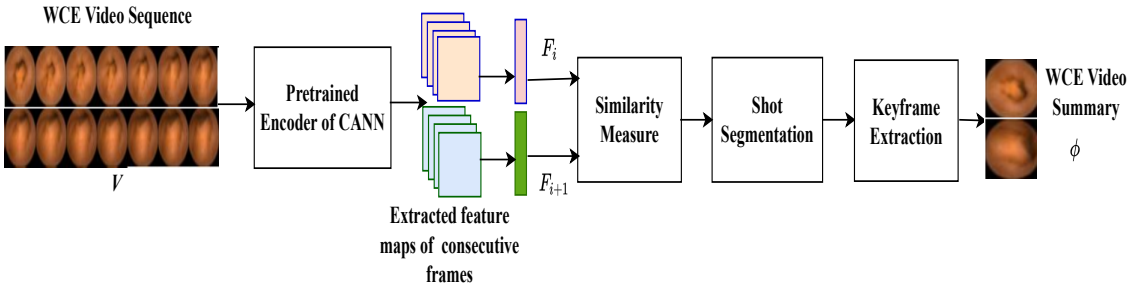


Figure 5.1: Proposed WCE video summarization framework; F_i and F_{i+1} are the feature vectors of i^{th} and $(i+1)^{th}$ sequential frames

WCE video sequence V consists of an ordered set of consecutive frames represented

as $V = I_1, I_2, I_3, ..., I_m$, where I_i denotes i^{th} frame of the video and m is the number of frames in video. Slow movement of the capsule in some parts of the GI tract such as small intestine results in small or no changes from frame to frame. In esophagus, WCE video exhibits large changes from frame to frame due to the fast movement of capsule. Therefore, a set of frames ϕ called as keyframes can be found which summarizes V by eliminating redundant frames. The task of finding ϕ given V involves the following function.

$$\left\{ I_V^{\phi_1}, I_V^{\phi_2}, ...I_V^{\phi_J} \right\} = \arg\min_{\phi_J} \left\{ D(\phi, V) | 1 \leq \phi_J \leq \kappa \right\}. \tag{5.1}$$

where D is measure of dissimilarity representing the criterion of video summarization. Proposed WCE-VS framework for constructing ϕ from V is shown in Figure 5.1.

5.3 CANN for Feature Extraction

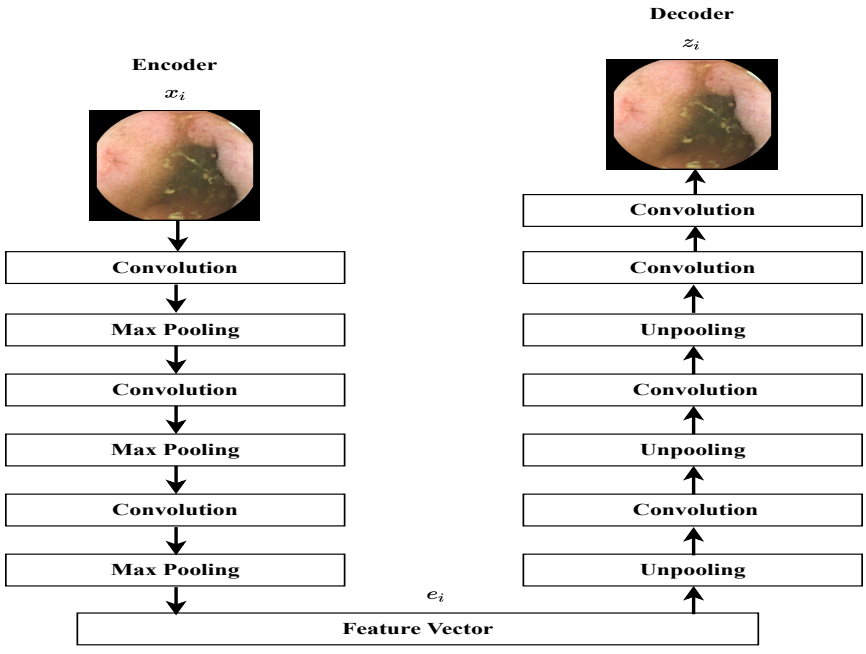


Figure 5.2: Convolutional autoencoder architecture showing encoder and decoder networks for extracting feature vector in endoscopic images

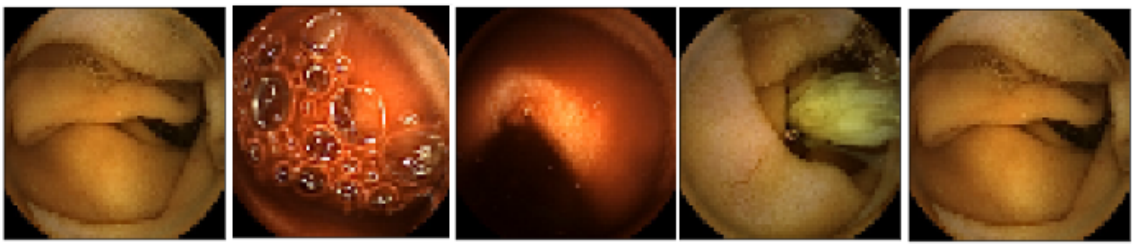
Convolutional autoencoder neural network (CANN) enables to extract features of endoscopic images using an unsupervised learning approach. Many research outcomes have shown that unsupervised feature extraction of medical images lead to significant improvement compared to conventional convolutional neural network (Kallen-

Table 5.1: Layer parameters of convolutional autoencoder

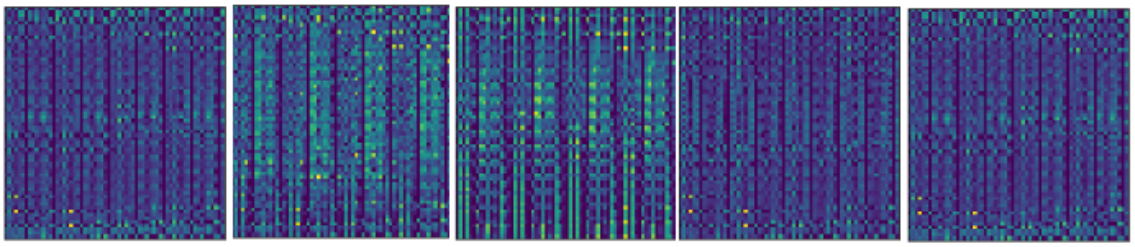
Layer	Type	Number of maps	Kernel Size	Output
1	Convolution	16	5x5	16x256x256
2	Max Pooling	16	2x2	16x128x128
3	Convolution	32	3x3	32x128x128
4	Max Pooling	32	2x2	32x64x64
5	Convolution	64	3x3	64x64x64
6	Max Pooling	64	2x2	64x32x32
7	Unpooling	64	2x2	64x64x64
8	Convolution	64	3x3	64x64x64
9	Unpooling	32	2x2	32x128x128
10	Convolution	32	3x3	32x128x128
11	Unpooling	16	2x2	16x256x256
12	Convolution	16	5x5	16x256x256
13	Convolution	3	3 x 3	3x256x256

berg *et al.* (2016), Kumar *et al.* (2015), Chen *et al.* (2017)). CANN consists of an encoder and decoder networks. Encoder of the CANN generates high level feature map of the input by using several convolution and max-pooling layers. Decoder reconstructs the input from the feature map by using unpooling and convolution layers. The proposed CANN is designed based on the autoencoder network described in (Masci *et al.* (2011)) and its architecture is shown in Figure 5.2. Proposed feature extraction approach utilizes convolutional filtering to train CANN in an unsupervised way.

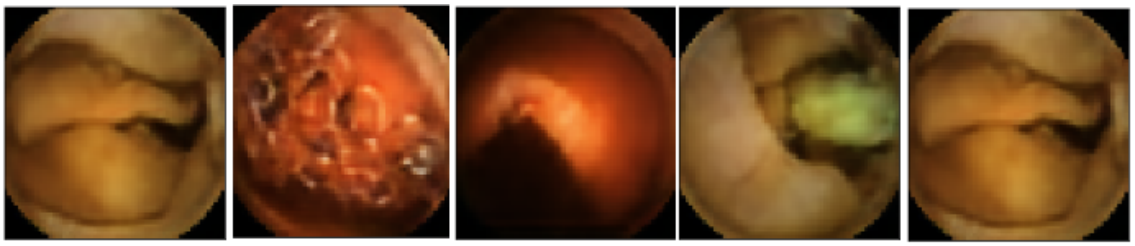
Encoder consists of three convolution layers and three max pooling layers. Max pooling layers down sample the feature maps extracted by convolution layers. Decoder consists stack of unpooling and convolution layers. Unpooling layers are used to upsample the feature maps extracted on the encoder side. Adding more layers will make the CANN more deeper and will improve the reconstruction of the input images at the decoder. But this increases the complexity of the model. CANN encoder with three convolution layers is capable of extracting the high level features from an image which can effectively discriminate the images into similar and dissimilar pairs. The final goal of the CANN is to find a feature vector for each input image by minimizing the mean squared error (MSE) between input and output image samples. The details of the encoder and decoder layer parameters are given in Table 5.1.



(a)



(b)



(c)

Figure 5.3: Visualization of extracted features and reconstructed images of CANN (a) Input images to CANN, (b) Features extracted by encoder network and (c) Decoded images by decoder network

Each convolution layer uses a non-linear activation function called Scaled Exponential Linear Unit (SELU) instead of Rectified Linear Units (ReLU) used in other convolutional networks. SELU activation function is close to zero mean and unit variance. When propagated through multiple network layers, SELU automatically converge towards zero mean and unit variance. All these self normalizing parameters of SELU makes learning highly robust in network with many layers and utilizes strong regularization schemes (Klambauer *et al.* (2017)). For an input image matrix x_i, the encoder network computes encoder output e_i using (5.2).

$$e_i = \sigma(x_i * f^n + b) \tag{5.2}$$

where σ denotes SELU activation function, $*$ represents 2D convolution operation, f^n is n^{th} convolutional filter kernel and b denotes encoder bias. The decoder reconstructs

the encoded output using (5.3).

$$z_i = \sigma(e_i * \tilde{f}^n + \tilde{b}) \tag{5.3}$$

where z_i is the reconstruction of the i^{th} input x_i, $\tilde{f}^n$ is the n^{th} decoder convolutional filter and $\tilde{b}$ is the bias of the decoder. Unsupervised training of the CANN aims to minimize the loss function given in (5.4).

$$J(\theta) = \sum_{i=1}^{m}(x_i - z_i)^2 \tag{5.4}$$

The gradients are computed using the loss function given in (5.4) and the network parameters are optimized through adam optimizer to minimize the reconstruction loss.

Similarity between the two consecutive frames is decided based on the features extracted from the frames. To extract the features of an input image in an unsupervised method, both the encoder and decoder networks are trained together. Input image is reconstructed by the decoder using encoder extracted features. Level of feature extraction is decided based on the reconstructed quality of images at the decoder. After the encoder is trained to extract the high level features, the decoder part of the CANN is removed and only the encoder is retained. Encoded features and reconstructed images of the CANN along with the input images are shown in Figure 5.3.

5.4 Similarity Estimation

Two consecutive frames in WCE video sequence is considered as an image pair. For any input image I_i to the CANN, the corresponding extracted feature Fea_i is generated as,

$$Fea_i = G\left(I_i, W\right) \tag{5.5}$$

where G(.) is a non-linear mapping function of the trained encoder and W is the network parameters of the encoder part of the CANN. SELU activation function is used as the non-linear activation function in convolution layers. Euclidean distance between features of the image pair is computed and classified as similar or dissimilar pair based on the fixed threshold. To learn the threshold, Euclidean distance between

6000 consecutive WCE image pairs in feature space is computed and the observations made are: i) Euclidean distance varies between around 0 to 270, ii) Similar pair of images have distance close to 0, iii) Dissimilar pair of images have larger distance close to 270, iv) Images with few dissimilar patches have distance approximately equal to 20. Based on all the above observations, a threshold of 20 is fixed for classification. Losing frames with significant lesions can be avoided by selecting small threshold. Euclidean distance for similarity judgement is calculated as:

$$Dis_{i,i+1} = ||(Fea_i - Fea_{i+1})| |_2 \tag{5.6}$$

where $Dis_{i,i+1}$ is the Euclidean distance of the features Fea_i and Fea_{i+1} extracted from i^{th} and $(i+1)^{th}$ WCE frames respectively. Based on $Dis_{i,i+1}$, the image pair is considered as similar or dissimilar by (5.7). Similar and dissimilar pair of images are labelled as 1 and 0 respectively.

$$S_i = \begin{cases} 1, & Dis_{i,i+1} < 20 \\ 0, & Dis_{i,i+1} \geq 20 \end{cases} \tag{5.7}$$

5.5 Shot Segmentation

A video shot is defined as group of contiguous frames segmented based on similarity changes between two consecutive frames. Proposed shot segmentation method is shown in Figure 5.4. Shot boundary is detected when the similarity label is set to 0. WCE shot segmentation which incorporates frame matching, separates shots consisting of frames with high similarity.

5.6 Keyframe Extraction

Video summarization of WCE video can be concluded with keyframe extraction in each shot. Frames in each shot has high similarity with a lot of redundancy. These redundant frames contributes very less or no information for covering the entire WCE video. Motion analysis between frames within a shot gives an idea of redundant frames (Zhu *et al.* (2005)). If a pair of frames exhibit larger motion then the frames are likely to be considered as less redundant. Inherent intra-shot redundancy is reduced to retrieve keyframe representation by analysing capsule's motion. Motion profile

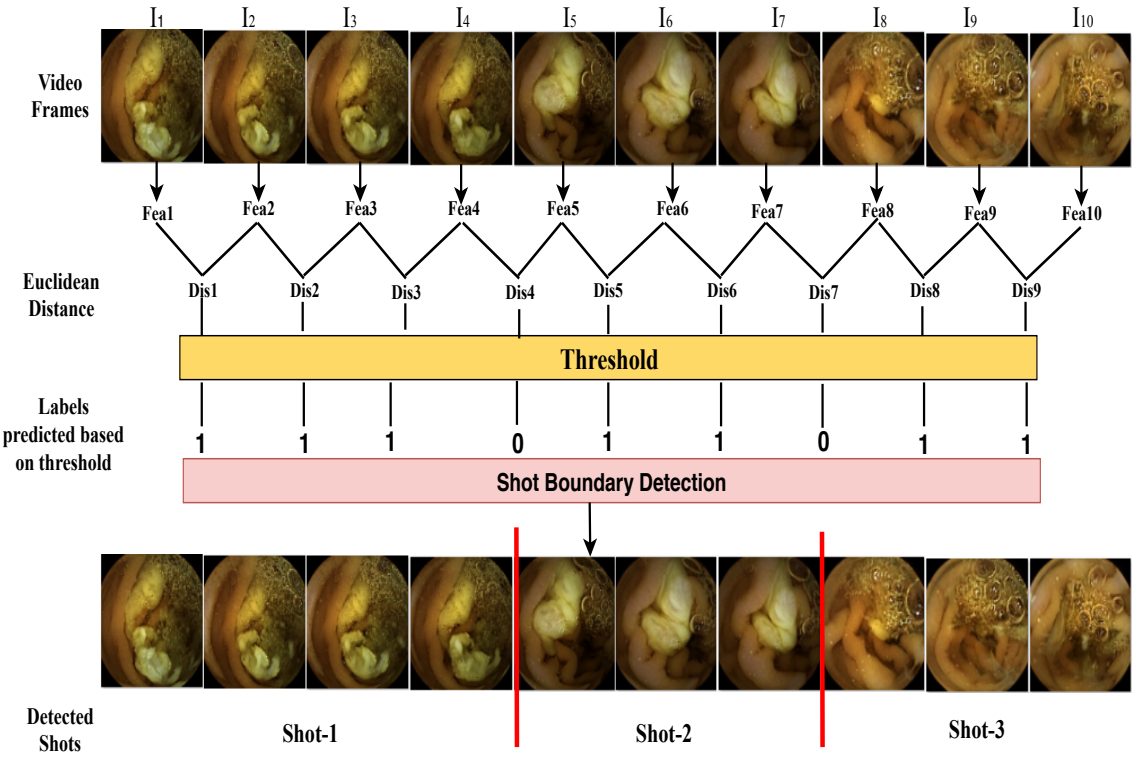


Figure 5.4: WCE video shot segmentation based on frame similarity

which constitutes motion score, motion direction and motion energy denoted as M_s, M_d, E_m respectively is derived from every shot before extracting keyframes.

First, the relative inter-frame motion score is estimated for a considered i^{th} shot S_{h_i} given as

$$M_{s_i} = \{M_{s_i}(n),\ n = 1, 2, ...n_i\}, \tag{5.8}$$

where n_i is the number of frames in S_{h_i} and $M_{s_i}(n)$ is intra-shot motion score between successive pair of frames.

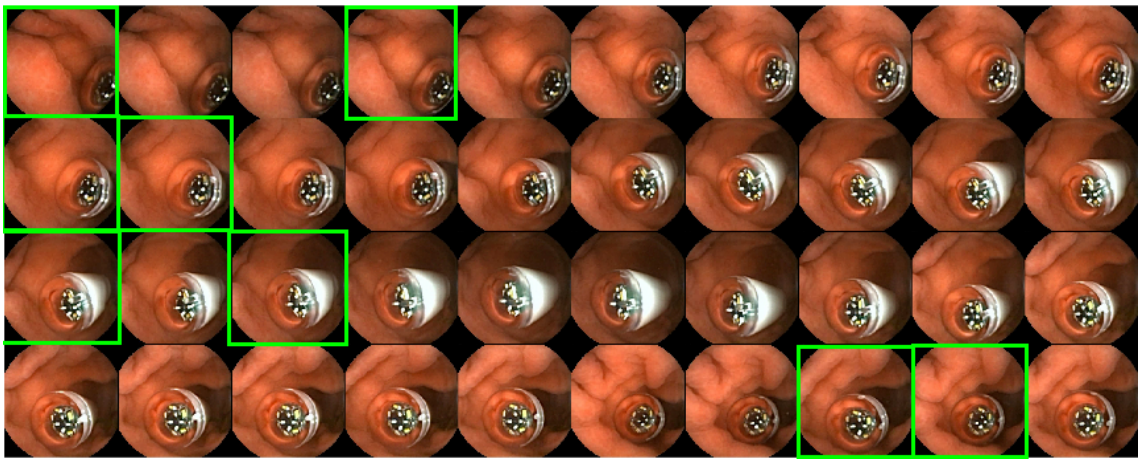
Motion score M_s for a frame pair (I, I') consisting of matched feature positions (X, X'), which is also the difference in average distance between (I, I') given by

$$M_s = \frac{1}{\alpha}\left\{\sum_{m=1}^{\alpha} d(x_m, \hat{x}_m) - \sum_{m=1}^{\alpha} d(x'_m, \hat{x}'_m)\right\} \tag{5.9}$$

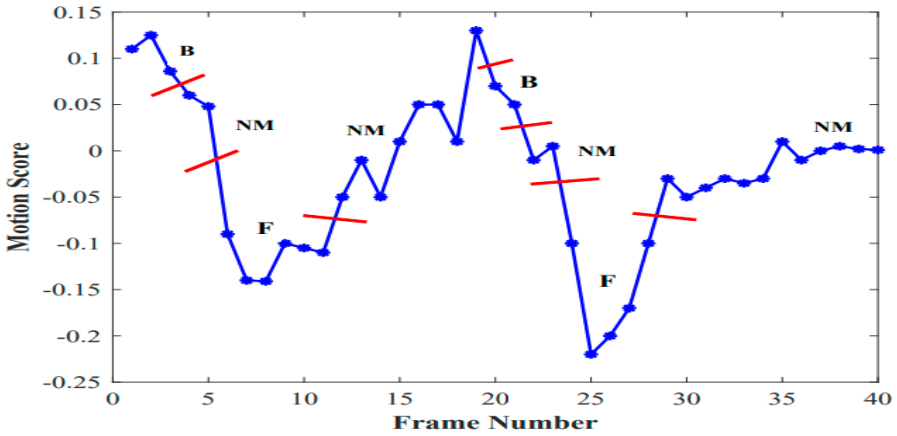
where $\hat{x}$ is the X features center of mass computed by:

$$\hat{x} = \frac{1}{\alpha}\sum_{m=1}^{\alpha} x_m \tag{5.10}$$

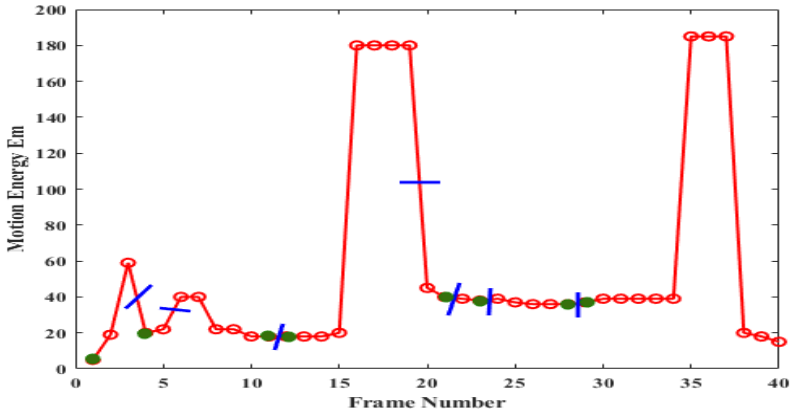
where α is number of matched feature pairs detected (Sargent *et al.* (2009)) and each (x_m, x'_m) is a matched feature pair in (X, X'). $d(x_m, \hat{x}_m)$ is the euclidean distance between x_m and $\hat{x}_m$.



(a)

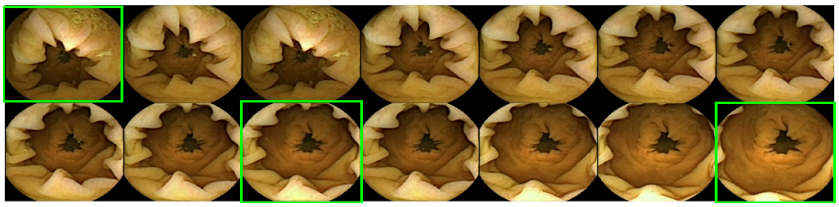


(b)

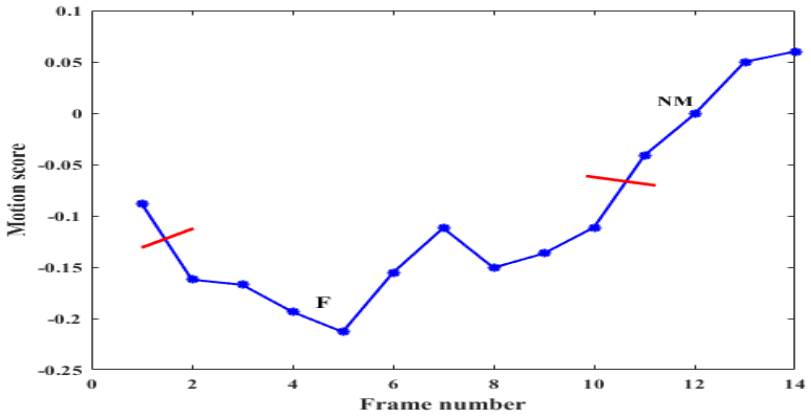


(c)

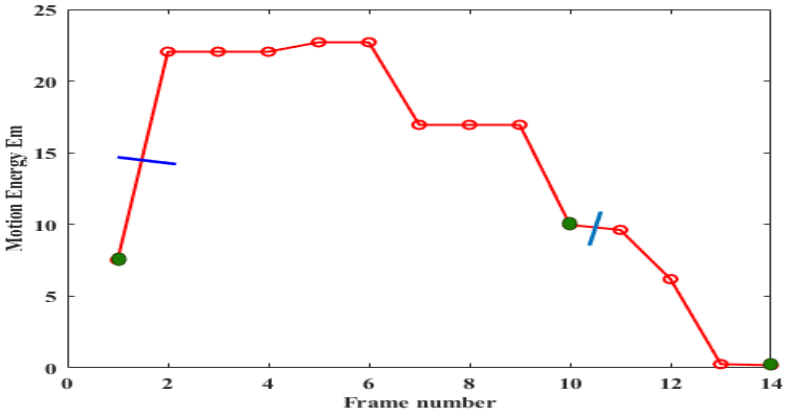
Figure 5.5: Keyframe selection of a 40 frame shot in video sequence captured in stomach based on motion profile. (a) Keyframes. (b) Motion signal partitioned into segments. (c) Motion energy signal.



(a)

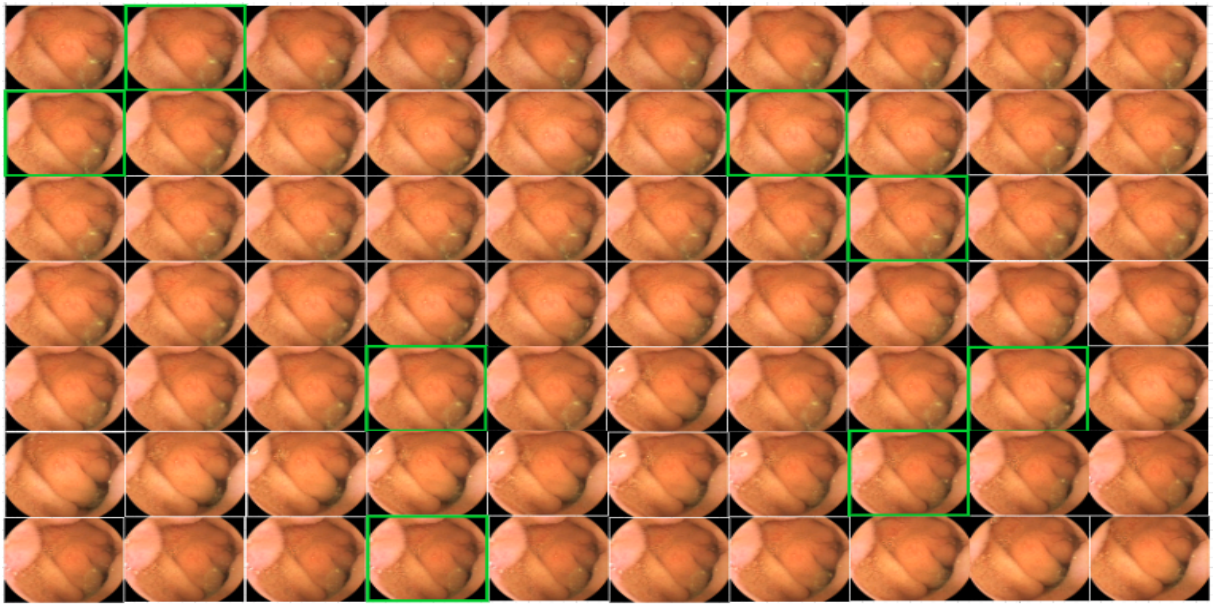


(b)

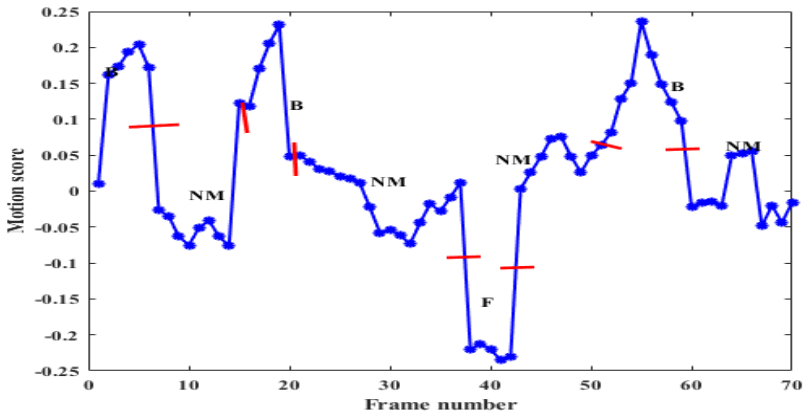


(c)

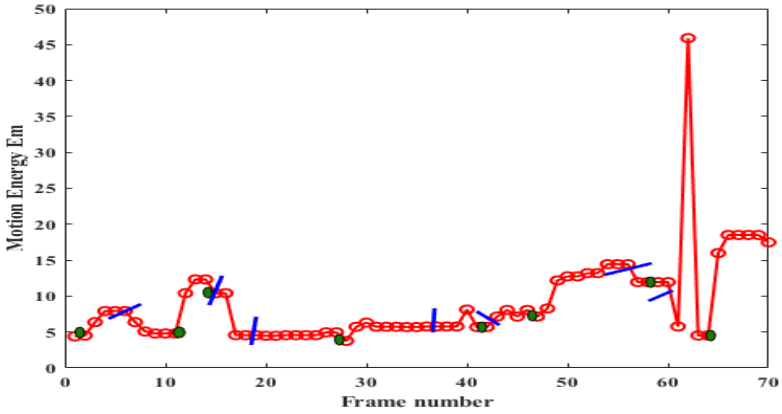
Figure 5.6: Keyframe selection of a 14 frame shot in colon video based on motion profile. (a) Keyframes. (b) Motion signal partitioned into segments. (c) Motion energy signal.



(a)



(b)



(c)

Figure 5.7: Visualization of a small bowel video shot with motion profile. (a) Keyframes (b) Partitioning of Motion signal. (b) Keyframe extraction based on motion energy

Next, the motion direction sequence $M_{d_i} = \{M_{d_i(n)},\ n = 1, 2, ...n_i\}$ is computed using (5.11), which classifies the motion direction as forward, backward and no-motion depending on the capsule's movement inside GI tract.

$$M_{d_i}(n) = \begin{cases} forward = 1, & \text{if } M_{s_i}(n) \leq TH_f \\ backward = -1, & \text{if } M_{s_i}(n) \geq TH_b \\ nomotion = 0 & otherwise \end{cases} \tag{5.11}$$

Considering a wide range of motion analysis through a large number of experiments, threshold values TH_f and TH_b are selected as -0.12 and 0.12. Finally, motion energy $E_{m_i} = \{E_{m_i}(n),\ n = 1, 2, ...n_i\}$ is computed for the features associated with each frame pair by:

$$E_m = \sum_{m=1}^{\alpha} ||x_m - x'_m||^2 \tag{5.12}$$

An example motion profile of a 40 frames in a shot from a video sequence captured in stomach in which the frame sequence exhibits forward, backward and no-motion is shown in Figure 5.5. Based on the obtained motion direction signal M_d of the capsule endoscope, an example shot is segmented into 8 different continuous runs consisting of forward, backward and no-motion indicated as F, B and NM respectively. Frame with minimum motion energy is selected as keyframe in each segment. Another motion profile of a short shot of 14 frames of video captured in colon which exhibits only forward and no-motion is shown in Figure 5.6. Keyframe selection in a video shot of 70 frames captured in a small bowel with motion profile is shown in Figure 5.7. In all the frame shots the keyframe indicators are marked by green circles.

5.7 Results & Discussions

5.7.1 Datasets

The proposed WCE-VS method is evaluated on two datasets given in Table 1.3 in Chapter 1. Around 5000 WCE frames resized to a resolution of 256 X 256, captured at different location of the GI tract from different patients are used for training CANN. Batchwise training is performed using mini-batch size of 8 samples and the number of epochs used for each batch is 50. For evaluating the performance of the proposed technique, keyframes of around 20 video sequences including 3 video sequences of the KID-dataset are identified with the help of an expert gastroenterlogist. Around 1000

similar and dissimilar pair of frames are identified to test the similarity judgement performance. These frame pairs and keyframes are used as ground-truth summary to compare the performance of the proposed technique with the other WCE VS methods. The details on frame resolution, frame motion characteristics and GI organ at which the frames are acquired are given in Table 1.3. Video sequences in KID dataset covers all the GI organs and exhibits organ dependent motion characteristics.

5.7.2 Performance Comparison

Proposed method is compared with the other unsupervised methods which involves different feature extraction, shot segmentation and key frame extraction methods. Methods with which the proposed method is compared are discussed below:

- Hue saturation value colour feature with K-means clustering (HSV-KMC) (Huo *et al.* (2012)): WCE images exhibits different mucosal feature characteristics. Color is one of the significant feature. GI organ vary in color features for different organs. These color features are extracted in Hue Saturation Value (HSV) color space, since the information associated with H component is more indicative in representing the differences in WCE images. Color feature vector is extracted by using histogram of H and S. Shot is detected when the consecutive pair of frames are having different color feature vector. In each shot the key frames are extracted by using K-means clustering (KMC) method.

- Colour, texture and shape features with K-means clustering (CTS-KMC) (Yuan and Meng (2013)): In this method, fusion of color, texture and shape (CTS) features are considered for shot detection. Color feature vector is created in HSV color space. Local binary pattern (LBP) algorithm is used to extract texture features. Shape features are represented using HoG. Entropy of extracted features for each frame is used for segmenting the video into different shots. KMC algorithm is used for key frames extraction.

- Scale-invariant feature transform with motion analysis (SIFT-MA) (Lowe (2004)): This method uses SIFT algorithm for feature extraction and matched feature points retrieved from two consecutive frames are used to detect a shot. Key frames are extracted based on motion analysis.

- Speeded up robust features with motion analysis (SURF-MA) (Bay *et al.* (2008)): This method uses SURF method for feature extraction and matched feature points from two consecutive frames to detect a shot. Key frames are extracted based on motion analysis.

Table 5.2: Comparison of Recall, Precision and F-score values of the proposed method with other methods on KID dataset

Parameters	Test Video	HSV-KMC	CTS-KMC	SIFT-MA	SURF-MA	Proposed Method
Recall	KID-1	0.57	0.79	0.80	0.78	0.94
	KID-2	0.54	0.74	0.79	0.73	0.92
	KID-3	0.51	0.72	0.83	0.76	0.93
	Average	**0.54**	**0.75**	**0.81**	**0.75**	**0.93**
Precision	KID-1	0.58	0.78	0.61	0.81	0.92
	KID-2	0.61	0.76	0.63	0.91	0.94
	KID-3	0.55	0.81	0.69	0.88	0.95
	Average	**0.58**	**0.78**	**0.64**	**0.86**	**0.94**
F-Score	KID-1	0.57	0.78	0.69	0.79	0.93
	KID-2	0.57	0.75	0.70	0.81	0.93
	KID-3	0.53	0.76	0.75	0.81	0.94
	Average	**0.56**	**0.76**	**0.72**	**0.80**	**0.93**

In the above SIFT and SURF based methods, matched feature points are retrieved between the pair of consecutive frames. The ratio of number of matched features to total number of features detected in both the frames is used to detect the shot. If the ratio is less than 0.15, it is considered that the two frames are in different shots. The comparison results for F-score on both the datasets are shown in Table 5.2 and Table 5.3. As seen from the table, the proposed method achieves better accuracy by 32%, 14%, 18% and 11% compared to HSV-KMC, CTS-KMC, SIFT-MA and SURF-MA methods respectively. The CR comparison results are given in Table 5.4. The proposed method provides the gain in CR by 10%, 9%, 8% and 7% compared to HSV-KMC, CTS-KMC, SIFT-MA and SURF-MA respectively. When compared to the benchmark methods, the proposed WCE-VS performs better in terms of both

Table 5.3: Comparison of Recall , Precision and F-score values of the proposed method with other methods on Dataset-2

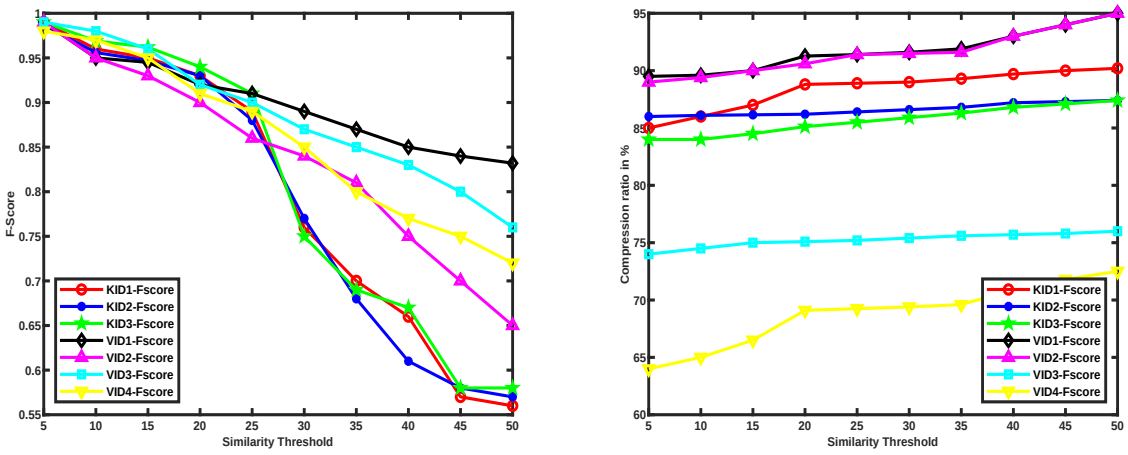
Parameters	Test Video	HSV-KMC	CTS-KMC	SIFT-MA	SURF-MA	Proposed method
Recall	Video-1	0.76	0.89	0.70	0.82	0.91
	Video-2	0.72	0.85	0.71	0.75	0.89
	Video-3	0.69	0.78	0.72	0.78	0.90
	Video-4	0.64	0.72	0.69	0.77	0.92
	Average	**0.70**	**0.81**	**0.70**	**0.78**	**0.90**
Precision	Video-1	0.72	0.84	0.94	0.89	0.94
	Video-2	0.58	0.79	0.90	0.88	0.92
	Video-3	0.51	0.79	0.82	0.82	0.95
	Video-4	0.53	0.81	0.81	0.89	0.91
	Average	**0.59**	**0.80**	**0.86**	**0.87**	**0.93**
F-Score	Video-1	0.74	0.86	0.80	0.85	0.92
	Video-2	0.64	0.82	0.79	0.80	0.90
	Video-3	0.58	0.78	0.76	0.80	0.92
	Video-4	0.59	0.76	0.74	0.82	0.91
	Average	**0.64**	**0.80**	**0.77**	**0.82**	**0.91**

Table 5.4: Comparison of the proposed method with other methods interms of F-score (FS) and compression ratio (CR) results in %

Test Video	HSV-KMC		CTS-KMC		SIFT-MA		SURF-MA		Proposed Method	
	FS	CR	FS	CR	FS	CR	FS	CR	FS	CR
KID-1	57.24	74.60	78.37	71.5	68.87	63.4	79.21	63.80	92.78	88.80
KID-2	56.98	72.30	74.72	68.6	70.04	69.57	81.37	72.20	93.06	86.20
KID-3	53.43	70.96	76.25	66.3	74.86	68.78	80.91	73.62	94.21	85.62
Video-1	74.16	89.50	86.19	86.2	80.12	85.31	84.79	85.27	91.91	91.27
Video-2	64.38	77.90	81.88	74.2	78.76	81.5	79.79	82.60	89.92	90.60
Video-3	57.78	75.7	78.36	73.3	76.21	79.26	80.03	81.09	92.06	75.09
Video-4	58.84	82.5	73.14	81.59	76.81	79.32	81.89	75.16	91.14	69.16
Average	**60.40**	**73.61**	**78.89**	**74.32**	**74.71**	**75.39**	**81.14**	**76.24**	**92.15**	**83.82**

F-score and CR.

It is very critical to achieve high accuracy in medical image analysis. The proposed method achieves high accuracy and high compression performance compared to other works. This indicates that it can eliminate redundant frames by extracting few key-frames which preserve informative frames.



(a) (b)

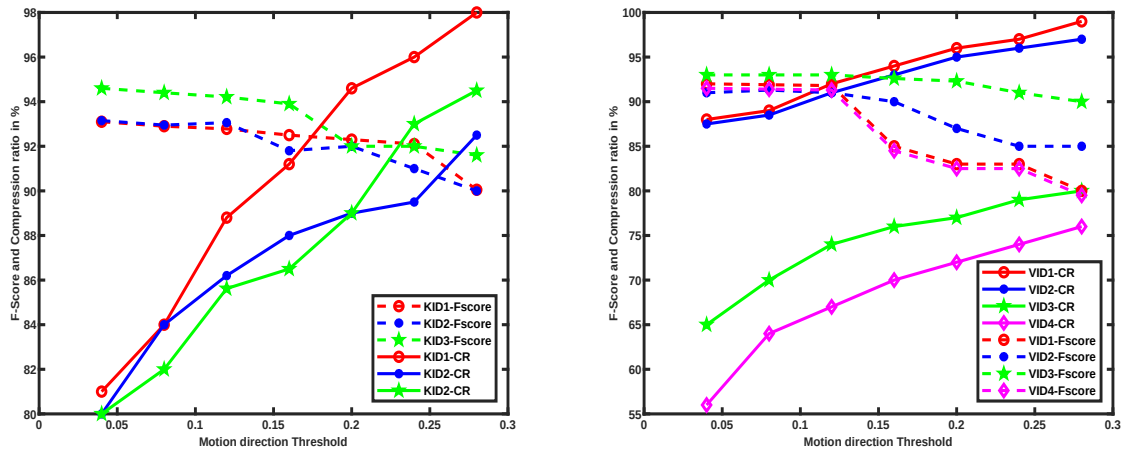
Figure 5.8: Performance test on similarity threshold. (a) F-score test on similarity threshold (b) Compression ratio test on similarity threshold

The frame similarity threshold used for shot detection has a direct impact on the F-score as shown in the Figure 5.8a. Each plot for a specific test video sequence indicates the performance measure variation according to the change in similarity threshold. Also it can be observed from Figure 5.8b that similarity threshold has less impact on the compression performance. Setting high similarity threshold can tend to reject frames with significant lesions. Therefore, it is necessary to choose low threshold value. In shot detection, a threshold of 20 is set for similarity estimation which detects even a small significant change between pair of frames and avoids loss of informative frames.

Video shot is partitioned into different motion segments based on a threshold and a keyframe is extracted in each segment. Construction of motion profile proved to be strong over thresholds TH_f and TH_b. Low threshold values -0.12 and 0.12 are chosen to detect even a weak motion between the frames, because TH_f and TH_b has direct impact on F-score and CR. F-score and CR for datasets considered in this work for

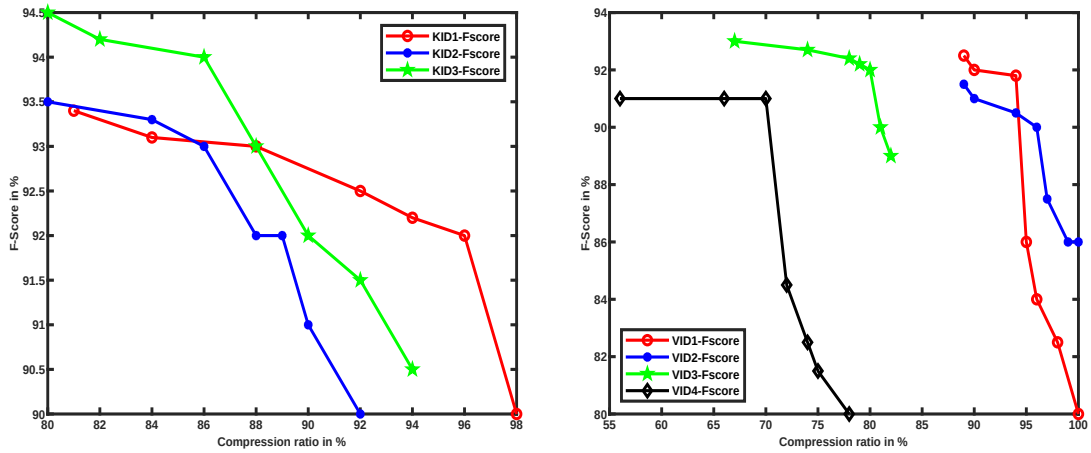
different motion direction thresholds are shown in Figure 5.9. From the performance graph, it can be observed that F-score is maximum at 0.12. Video shot is partitioned into less number of segments at higher threshold and less keyframes are selected. Therefore, CR increases for the larger thresholds. But accuracy is very important and threshold at which high accuracy is achieved is considered in the work.

As shown in Figure 5.10, the summarization performance in-terms of F-score decreases as CR increases. Each plot for a particular test video sequence indicates how the F-score varies as CR varies. It can be observed from Figure 5.10b that video with slow motion (Video-1 in dataset-2) has high F-score of around 93% at high CR of 95%. Video with fast motion (Video-4 in dataset-2) achieves high F-score of 91% with 70% CR. The accuracy drastically drops as the CR increases which is directly influenced by increase in TH_f and TH_b. Larger TH_f and TH_b gives high CR as more number of frame pairs are considered as no motion frames and this results in less motion segments. This will lead to an excessive rejection of the significant frames and affects summarization performance interms of accuracy. The results clearly indicate that the proposed method is potential with consistent performance with greater than 90% accuracy achieved for video sequences of different motion characteristics.



(a) (b)

Figure 5.9: F-score and Compression performance for different motion direction thresholds . (a) KID-dataset (b) Dataset-2.



(a) (b)

Figure 5.10: Comparison of summarization performance in-terms of F-score with compression ratio on (a) KID-dataset (b) Dataset-2.

5.8 Summary

A framework to obtain summary of WCE video content is presented. Convolutional autoencoder is trained to extract high level features in an unsupervised way, which are used to estimate the similarity between the consecutive frames. Based on the similarity measure the video is segmented into different shots. The unsupervised training method avoids laborious procedure of labelling large number of WCE image pairs to detect video shots. The change in two successive frames of WCE video is due to capsule motion, which varies in different parts of GI tract. Therefore, the keyframes are extracted based on the motion profile constructed for each video shot. This method eliminates frames with slight temporal differences and retains candidate keyframes covering sufficient WCE video. Similarity and motion detection thresholds have a key role in deciding the summarization performance interms of F-Score and compression ratio. Thresholds are set to get maximum accuracy in this work. With the set thresholds, the proposed method achieves an average F-measure of 91.1% with compression ratio of 83.12%.

Chapter 6

Conclusions and Future Directions

6.1 Conclusion

Although WCE is the most preferred modality for diagnosing and assessing small intestinal disorders, the diagnostic yield is limited due to low video resolution. Enhancement in resolution improves the diagnostic performance but increases the video data for processing and transmission. An increase in video data consumes more processing and transmission power, which is not feasible due to limited capsule battery power. Video data can be substantially reduced by using a low complexity video compression technique that does not consume much power. In this research work, a low complexity video encoder of the wireless capsule endoscope with a decoder for reconstructing the compressed video at acceptable medical image quality is proposed. The WCE procedure captures a video with a huge number of frames, which is considered for review by a physician after completing the entire procedure. Reviewing a large number of frames at once is a tedious task that requires attention and expertise. Moreover, most of the frames in the video are redundant, which can be removed by extracting only the keyframes. A video summarization framework to generate a summary consisting only keyframes is proposed in this research work.

A low complexity DVC-FBC architecture is proposed in Chapter 3 for WCE video compression. The complexity reduction in keyframe encoding is achieved by exploiting GI image textural characteristics. Further, modifications have been made in the transform and quantization functional stages to reduce the computations. The quality of SI creation at the decoder determines the compression performance of DVC. This is accomplished by generating good SI quality at the decoder by employing intra coded

low frequency components as the hash. The quality of SI goes upto 36 dB and SI refinement can be done with few transmitted parity bits which improves the compression performance. SI generation depends only on the previous encoded frames which enables increase in GOP size and reduces the number of keyframes. Therefore, more number of WZ frames can exploit the temporal correlation to improve the compression performance. Latency in SI generation is reduced as the process depends only on available reconstructed frames. A new approach for encoding of chroma component of WZ frame and SI generation for chroma is presented. The assessment of the proposed DVC system is done by using rate-distortion performance and encoding complexity. Better performance is achieved compared to MJPEG by 60% of BD-bitrate savings with PSNR gain of 6 dB at increased encoding complexity. Compared to TDWZ based DVC, the proposed achieved 40% bitrate savings with 5 dB PSNR gain at reduced encoder complexity. Though, H.264-Intra performs better than proposed in terms of RD, its complexity is 3 - 4 times higher than the the proposed.

In Chapter 4, an improvised version of DVC-FBC referred to as DVC-DCP is presented which achieves better RD performance with reduced encoder complexity. Reduced encoder complexity is achieved by eliminating WZ frames chroma processing at the encoder. A deep CNN model is trained to predict chroma of the WZ frame at the decoder by matching luma and texture information of the keyframe and WZ frames. The deep chroma prediction model comprises a merging block with a spatial attention mechanism to make use of spatial inter frame correlation for accurate matching of similar regions for transferring chroma. The performance of the deep chroma prediction model is evaluated by colour similarity and quality metrics. The model used at the decoder to predict chroma using keyframe and WZ-luma performs better compared to chroma reconstruction obtained by dequantization and upsampling. When the video is encoded at GOP=4, 25% of the frames are keyframe encoded and while the remaining are WZ encoded. Encoding complexity of chroma component of 75% of the frames is reduced at the improved compression efficiency. DVC-DCP model achieved an RD performance close to H.264-Intra with a much lesser encoding complexity comparable with that of MJPEG. DVC-DCP achieves BD-bitrate savings of 65%, 54% and 25% with PSNR gain 10 dB, 7.16 dB and 2.43 dB for MJPEG, TDWZ-DVC and DVC-FBC respectively. H.264-Intra performs better in RD performance,

but the encoder complexity is 3 - 4 times greater than DVC-DCP. The simulation results justify that the proposed DVC-DCP achieves improved RD performance with reduced encoding complexity.

Manual reviewing of the huge amount of frames captured during the WCE procedure is challenging for the physician in terms of time and accurate diagnosis. To overcome this, a computer-aided WCE video summarization framework consisting convolutional autoencoder to extract deep features is presented in Chapter 5. A video is segmented into shots based on the similarity between the extracted deep features. Deep features provide better segmentation compared to handcrafted features. Keyframes from each shot are extracted based on the motion profile created using motion energy, motion direction and motion score between two consecutive frames. The proposed method achieves an average F-measure of 91.1% with compression ratio of 83.12%. Reduction in 83% of the total frames results in fewer number of frames for the review and hence improves diagnostic performance.

6.2 Future Directions

WCE is the most preferred procedure for diagnosis and analysis of GI abnormalities. The major drawback of this procedure is the poor image resolution which limits both the manual and computer-aided diagnosis techniques. Many studies have found that the high resolution medical imagery provides better diagnostic performance. In WCE, improved image resolution can be achieved by increasing the area of the lens and camera sensor array, however this is not always a viable option for many endoscopic applications due to computational cost and resource constraints. The computational cost can be reduced upto some extent by using high performance low complexity encoder techniques. To solve this problem, the computer vision research has developed a set of super-resolution techniques, which are used to generate high-resolution images from low-resolution imaging devices. High resolution images can improve the identification and locating of abnormalities in the images. Therefore, methods to generate high resolution video from low resolution WCE video can be considered as future work.

When the capsule travels in the GI tract, the frames captured are sometimes com-

pletely or partially degraded due to poor illumination and obscured by secreted fluids. These frames are either completely uninformative or partly informative. Methods to detect and remove the uninformative frames to reduce the video content for the review and restoration techniques for improving the quality of partially degraded frames can be considered as future work.

Appendices

A Distributed Coding of Correlated Frames

In conventional or predictive video encoding the probabilistic correlation between two frames F1 and F2 is available to both encoder and decoder. As per Shannon's source coding theorem lowest rate bound achievable for lossless compression is joint entropy H(F1,F2) of F1 and F2. In distributed video coding, the two frames are encoded independently and joint decoded by exploiting the inter-frame correlation.

A.1 Slepian-Wolf Coding

According to Slepian-Wolf (SW) theorem, two frames called as keyframe and WZ frame are encoded by different encoders and decoded by a combination of two decoders to reconstruct the frames. For a lossless compression, the achievable rate for decoding F1 and F2 with small probability of error is given by:

$$R_{F1} \geq H(\frac{F1}{F2})$$
$$R_{F2} \geq H(\frac{F2}{F1})$$
$$R_{F1} + R_{F2} \geq H(F2, F1)$$

Where, $H(\frac{F2}{F1})$ and $H(\frac{F1}{F2})$ are the conditional entropies.

In parity-bits generation approach of SW encoding, parity-check bits of a systematic code are employed. In order to encode n-bits, (n+r,n) systematic channel code defined by generator matrix $G_{nX(n+r)} = [I_n|P_{nXr}]$ is used. Compressed data represented in the form of parity bits of length r-bits are computed at the encoder by using parity matrix P and transmitted to the decoder. At the decoder these parity bits are concatenated to the n-bits generated from the side-information to form a bit-plane of length n+r. The decoded codeword is produced by parity-based SW decoder $G'_{nX(n+r)}$. Lossy compression achieved by quantization and SW encoding is called as Wyner-Ziv coding.

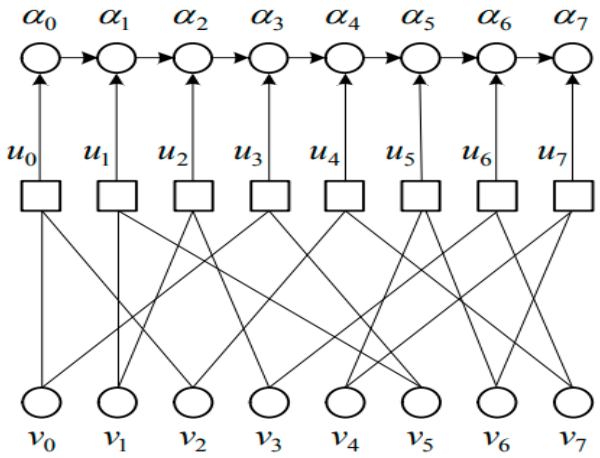


Figure A1: LDPCA encoding structure (Varodayan *et al.* (2011))

A.2 Low Density Parity Check (LDPC) codes in SW coding

The encoder transmits a weak channel code initially, and the decoder performs decoding using approximate channel statistics in the rate-adaptive LDPC coding method. If decoding is successful, the decoder instructs the encoder to move on to a new block. If decoding is unsuccessful, the encoder creates a lengthier syndrome depending on a lower-rate code to complement the broadcast channel code. This process is repeated until the syndrome meets the criteria for effective decoding. This strategy is appropriate under the conditions of feedback channel availability and low-delay requirements.

LDPC accumulate (LDPCA) codes are created by concatenating an LDPC syndrome with an accumulator. The LDPCA encoder generates syndrome bits $\mathbf{s}$ by applying mod-2 addition of the source bits $\mathbf{x}$ according to the LDPC factor graph, just like in traditional syndrome-based LDPC schemes. Unlike other procedures, the resultant syndrome bits are then mod-2 accumulated to produce the accumulated syndrome bits. This operation is depicted in Figure. A1 for illustrative purposes. The encoder keeps track of the accumulated syndrome bits in buffer and sends them to the decoder when requested.

The LDPCA decoder may accomplish a variety of rates by adjusting the LDPC decoding factor graph in response to the receipt of a new increment of the accumulated syndrome. As a result, the graph resulting from the various punctured syndrome patterns retains the degree of all variable nodes as described by the non-punctured LDPC code's parity-check matrix shown in Figure. A2. As a result, this approach allows for more soft information to be exchanged between the variable and the punc-

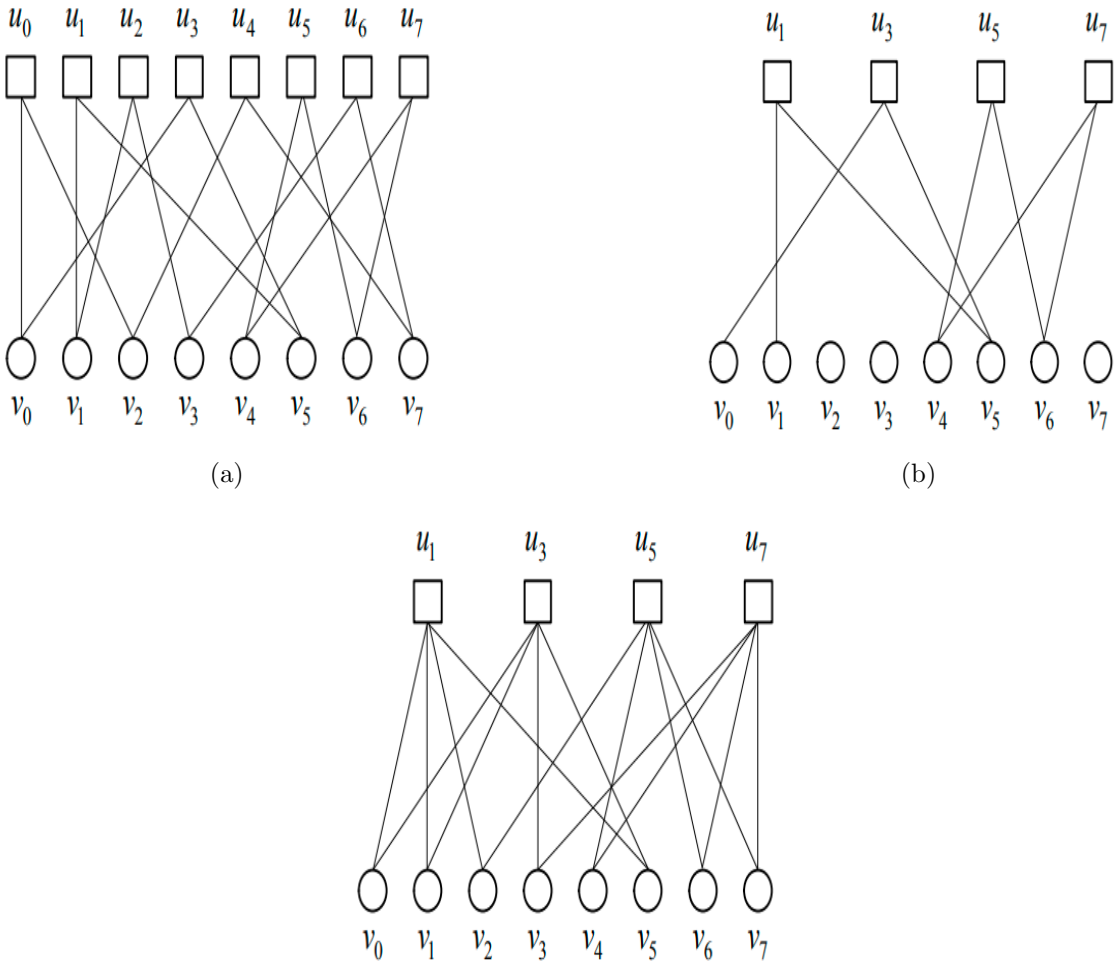


Figure A2: LDPC decoding structure (Varodayan *et al.* (2011)). (a) Entire LDPC structure, (b) Resultant LDPC structure for even indexed syndrome bits, (c) Resulting structure for even indexed accumulated syndrome bits

tured check nodes, resulting in a more effective iterative decoding than traditional punctured LDPC codes. To conclude, the complexity of encoding and decoding of LDPCA code depends on the length of bit-plane.

References

Aaron, A., **S. Rane**, and **B. Girod**, Wyner-Ziv video coding with hash-based motion compensation at the receiver. *In 2004 International Conference on Image Processing, 2004. ICIP'04.*, volume 5. IEEE, 2004.

Alam, M. W., **M. M. Hasan**, **S. K. Mohammed**, **F. Deeba**, and **K. A. Wahid** (2017). Are current advances of compression algorithms for capsule endoscopy enough? a technical review. *IEEE reviews in biomedical engineering*, **10**, 26–43.

Andreuccetti, D. (2012). An internet resource for the calculation of the dielectric properties of body tissues in the frequency range 10 hz-100 ghz. *http://niremf. ifac. cnr. it/tissprop/*.

Apostolidis, E., **E. Adamantidou**, **A. I. Metsai**, **V. Mezaris**, and **I. Patras** (2021). Video summarization using deep neural networks: A survey. *arXiv preprint arXiv:2101.06072*.

Artigas, X., **J. Ascenso**, **M. Dalai**, **S. Klomp**, **D. Kubasov**, and **M. Ouaret**, The DISCOVER codec: architecture, techniques and evaluation. *In picture coding symposium (PCS'07)*, CONF. 2007.

Ascenso, J., **C. Brites**, and **F. Pereira** (2010). A flexible side information generation framework for distributed video coding. *Multimedia Tools and Applications*, **48**(3).

Avni, D., **G. Meron**, **E. Horn**, **O. Zinaty**, and **A. Glukhovsky** (2010). Diagnostic device, system and method for reduced data transmission. US Patent 7,664,174.

Badrinarayanan, V., **A. Kendall**, and **R. Cipolla** (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, **39**(12), 2481–2495.

Bandy, W. R., **B. G. Jamieson**, **K. J. Powell**, **K. E. Salsman**, **R. C. Schober**, **J. Weitzner**, and **M. R. Arneson** (2013). Ingestible endoscopic optical scanning device. US Patent 8,529,441.

Barducci, L., **J. C. Norton**, **S. Sarker**, **S. Mohammed**, **R. Jones**, **P. Valdastri**, and **B. S. Terry** (2020). Fundamentals of the gut for capsule engineers. *Progress in Biomedical Engineering*, **2**(4), 042002.

Bay, H., **A. Ess**, **T. Tuytelaars**, and **L. Van Gool** (2008). Speeded-up robust features (SURF). *Computer vision and image understanding*, **110**(3), 346–359.

Biniaz, A., **R. A. Zoroofi**, and **M. R. Sohrabi** (2020). Automatic reduction of wireless capsule endoscopy reviewing time based on factorization analysis. *Biomedical Signal Processing and Control*, **59**, 101897.

Bjontegaard, G. (2001). Calculation of average PSNR differences between RD-curves. *VCEG-M33*.

Borchert, S., **R. Westerlaken**, **R. K. Gunnewiek**, and **R. Lagendijk** (2007). On extrapolating side information in distributed video coding. *26th Picture Coding System*.

Boudechiche, D. E., **S. Benierbah**, and **M. Khamadja** (2017). Distributed video coding based on vector quantization: application to capsule endoscopy. *Journal of Visual Communication and Image Representation*, **49**, 14–26.

Brack, T., **M. Alles**, **T. Lehnigk-Emden**, **F. Kienle**, **N. Wehn**, **N. E. L'Insalata**, **F. Rossi**, **M. Rovini**, and **L. Fanucci**, Low complexity LDPC code decoders for next geenration standards. *In 2007 Design, Automation & Test in Europe Conference & Exhibition*. IEEE, 2007.

Brites, C. and **F. Pereira** (2008). Correlation noise modeling for efficient pixel and transform domain wyner–ziv video coding. *IEEE Transactions on Circuits and systems for Video Technology*, **18**(9), 1177–1190.

Chen, J., **Y. Wang**, and **Y. Zou**, An adaptive redundant image elimination for wireless capsule endoscopy review based on temporal correlation and color-texture feature similarity. *In 2015 IEEE International Conference on Digital Signal Processing (DSP)*. IEEE, 2015.

Chen, J., **Y. Zou**, and **Y. Wang**, Wireless capsule endoscopy video summarization: a learning approach based on siamese neural network and support vector machine. *In 2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016.

Chen, M., **X. Shi**, **Y. Zhang**, **D. Wu**, and **M. Guizani** (2017). Deep features learning for medical image analysis with convolutional autoencoder neural network. *IEEE Transactions on Big Data*.

Chen, X., **X. Zhang**, **L. Zhang**, **X. Li**, **N. Qi**, **H. Jiang**, and **Z. Wang** (2009). A wireless capsule endoscope system with low-power controlling and processing ASIC. *IEEE Transactions on Biomedical Circuits and Systems*, **3**(1), 11–22.

Chen, Y., **Y. Lan**, and **H. Ren**, Trimming the wireless capsule endoscopic video by removing redundant frames. *In 2012 8th International Conference on Wireless Communications, Networking and Mobile Computing*. IEEE, 2012.

Chenb, X., **X. Zhang**, **L. Zhang**, **X. Li**, **N. Qi**, **H. Jiang**, and **Z. Wang** (2009). A wireless capsule endoscope system with low-power controlling and processing asic. *IEEE Transactions on Biomedical Circuits and Systems*, **3**(1), 11–22.

Ciuti, G., **A. Menciassi**, and **P. Dario** (2011). Capsule endoscopy: From current achievements to open challenges. *IEEE reviews in biomedical engineering*, **4**, 59–72.

Dalal, N. and **B. Triggs**, Histograms of oriented gradients for human detection. *In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1. IEEE, 2005.

Deligiannis, N., **J. Barbarien**, **M. Jacobs**, **A. Munteanu**, **A. Skodras**, and **P. Schelkens** (2012*a*). Side-information-dependent correlation channel estimation in hash-based distributed video coding. *IEEE Transactions on Image Processing*, **21**(4).

Deligiannis, N., **F. Verbist**, **J. Barbarien**, **J. Slowack**, **R. Van de Walle**, **P. Schelkens**, and **A. Munteanu**, Distributed coding of endoscopic video. *In 2011 18th IEEE International Conference on Image Processing*. IEEE, 2011.

Deligiannis, N., **F. Verbist**, **A. C. Iossifides**, **J. Slowack**, **R. Van de Walle**, **P. Schelkens**, and **A. Munteanu** (2012*b*). Wyner-Ziv video coding for wireless lightweight multimedia applications. *EURASIP Journal on Wireless Communica-*

tions and Networking, **2012**(1), 106.

Dung, L.-R., **Y.-Y. Wu**, **H.-C. Lai**, and **P.-K. Weng**, A modified H. 264 intra-frame video encoder for capsule endoscope. *In 2008 IEEE Biomedical Circuits and Systems Conference*. IEEE, 2008.

Fante, K. A., **B. Bhaumik**, and **S. Chatterjee** (2016). Design and implementation of computationally efficient image compressor for wireless capsule endoscopy. *Circuits, Systems, and Signal Processing*, **35**(5), 1677–1703.

Freedman, D. and **P. Kisilev**, Object-to-object color transfer: Optimal flows and smsp transformations. *In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010.

Glukhovsky, A., **D. Avni**, and **G. Meron** (2003). Diagnostic device using data compression. US Patent App. 10/202,626.

Gordon, S., **D. Marpe**, and **T. Wiegand** (2004). Simplified use of 8×8 transforms–updated proposal and results. *Joint Video Team(JVT) of ISO/IEC MPEG and ITU-T VCEG*.

Gowda, S. N. and **C. Yuan**, Colornet: Investigating the importance of color spaces for image classification. *In Asian Conference on Computer Vision*. Springer, 2018.

Gu, Y., **X. Xie**, **G. Li**, **T. Sun**, and **Z. Wang** (2012). Two-stage wireless capsule image compression with low complexity and high quality. *Electronics letters*, **48**(25), 1588–1589.

Guo, Z., **L. Zhang**, and **D. Zhang** (2010). A completed modeling of local binary pattern operator for texture classification. *IEEE transactions on image processing*, **19**(6), 1657–1663.

Gurudu, S. R., **H. E. Vargas**, and **J. A. Leighton** (2008). New frontiers in small-bowel imaging: the expanding technology of capsule endoscopy and its impact in clinical gastroenterology. *Reviews in gastroenterological disorders*, **8**(1), 1.

Gygli, M., **H. Grabner**, **H. Riemenschneider**, and **L. Van Gool**, Creating summaries from user videos. *In European conference on computer vision*. Springer, 2014.

Hamzaoglu, I., **O. Tasdizen**, and **E. Sahin** (2008). An efficient H. 264 intra frame coder system. *IEEE Transactions on Consumer Electronics*, **54**(4), 1903–1911.

He, J.-Y., **X. Wu**, **Y.-G. Jiang**, **Q. Peng**, and **R. Jain** (2018). Hookworm detection in wireless capsule endoscopy images with deep learning. *IEEE Transactions on Image Processing*, **27**(5), 2379–2392.

Hernandez-Lara, A. and **E. Rajan** (2021). Training, reading, and reporting for small bowel video capsule endoscopy. *Gastrointestinal Endoscopy Clinics*, **31**(2), 237–249.

Horn, E. (2008). Device, system, and method for reducing image data captured in-vivo. US Patent 7,336,833.

Huang, J., **T. N. Kumar**, **H. A. Almurib**, and **F. Lombardi** (2019). A deterministic low-complexity approximate (multiplier-less) technique for DCT computation. *IEEE Transactions on Circuits and Systems I: Regular Papers*.

Huo, J. S., **Y. X. Zou**, and **L. Li**, An advanced WCE video summary using relation matrix rank. *In Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics*. IEEE, 2012.

Iakovidis, D. K., **S. Tsevas**, and **A. Polydorou** (2010). Reduction of capsule endoscopy reading times by unsupervised image mining. *Computerized Medical Imaging and Graphics*, **34**(6), 471–478.

Iddan, G., **G. Meron**, **A. Glukhovsky**, and **P. Swain** (2000). Wireless capsule endoscopy. *Nature*, **405**(6785), 417.

Ishikura, K., **N. Kurita**, **D. M. Chandler**, and **G. Ohashi** (2017). Saliency detection based on multiscale extrema of local perceptual color differences. *IEEE Transactions on Image Processing*, **27**(2), 703–717.

Ismail, M. M. B., **O. Bchir**, and **A. Z. Emam**, Endoscopy video summarization based on unsupervised learning and feature discrimination. *In 2013 Visual Communications and Image Processing (VCIP)*. IEEE, 2013.

Istepanian, R., **N. Philip**, **M. Martini**, **N. Amso**, and **P. Shorvon**, Subjective and objective quality assessment in wireless teleultrasonography imaging. *In 2008*

30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2008.

Jia, X., **X. Xing**, **Y. Yuan**, **L. Xing**, and **M. Q.-H. Meng** (2019). Wireless capsule endoscopy: A new tool for cancer screening in the colon with deep-learning-based polyp recognition. *Proceedings of the IEEE*, **108**(1), 178–197.

Kallenberg, M., **K. Petersen**, **M. Nielsen**, **A. Y. Ng**, **P. Diao**, **C. Igel**, **C. M. Vachon**, **K. Holland**, **R. R. Winkel**, **N. Karssemeijer**, *et al.* (2016). Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE transactions on medical imaging*, **35**(5), 1322–1331.

Kanungo, T., **D. M. Mount**, **N. S. Netanyahu**, **C. D. Piatko**, **R. Silverman**, and **A. Y. Wu** (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, **24**(7), 881–892.

Khan, T., **R. Shrestha**, **M. S. Imtiaz**, and **K. A. Wahid** (2015). Colour-reproduction algorithm for transmitting variable video frames and its application to capsule endoscopy. *Healthcare technology letters*, **2**(2), 52–57.

Khan, T. H., **S. K. Mohammed**, **M. S. Imtiaz**, and **K. A. Wahid** (2016). Color reproduction and processing algorithm based on real-time mapping for endoscopic images. *SpringerPlus*, **5**(1), 1–16.

Khan, T. H. and **K. A. Wahid** (2011*a*). Lossless and low-power image compressor for wireless capsule endoscopy. *VLSI design*, **2011**.

Khan, T. H. and **K. A. Wahid** (2011*b*). Low power and low complexity compressor for video capsule endoscopy. *IEEE Transactions on Circuits and Systems for Video Technology*, **21**(10), 1534–1546.

KID Dataset (2017). Available online: https://mdss.uth.gr/datasets/endoscopy/kid/.

Klambauer, G., **T. Unterthiner**, **A. Mayr**, and **S. Hochreiter**, Self-normalizing neural networks. *In Advances in neural information processing systems*. 2017.

Klang, E., **Y. Barash**, **R. Y. Margalit**, **S. Soffer**, **O. Shimon**, **A. Albshesh**, **S. Ben-Horin**, **M. M. Amitai**, **R. Eliakim**, and **U. Kopylov** (2020). Deep

learning algorithms for automated detection of Crohns disease ulcers by video capsule endoscopy. *Gastrointestinal endoscopy*, **91**(3), 606–613.

Kumar, D., **A. Wong**, and **D. A. Clausi**, Lung nodule classification using deep features in CT images. *In 2015 12th Conference on Computer and Robot Vision*. IEEE, 2015.

Lee, H.-G., **M.-K. Choi**, **B.-S. Shin**, and **S.-C. Lee** (2013). Reducing redundancy in wireless capsule endoscopy videos. *Computers in Biology and Medicine*, **43**(6), 670–682.

Li, B. and **M. Q.-H. Meng** (2009). Computer-aided detection of bleeding regions for capsule endoscopy images. *IEEE Transactions on biomedical engineering*, **56**(4), 1032–1039.

Li, B., **M. Q.-H. Meng**, and **Q. Zhao**, Wireless capsule endoscopy video summary. *In 2010 IEEE International Conference on Robotics and Biomimetics*. IEEE, 2010.

Li, B. N. N., **X. Wang**, **R. Wang**, **T. Zhou**, **R. Gao**, **E. J. Ciaccio**, and **P. H. Green** (2019). Celiac disease detection from videocapsule endoscopy images using strip principal component analysis. *IEEE/ACM transactions on computational biology and bioinformatics*.

Lin, M.-C. and **L.-R. Dung** (2011*a*). A subsample-based low-power image compressor for capsule gastrointestinal endoscopy. *EURASIP Journal on Advances in Signal Processing*, **2011**, 1–15.

Lin, M.-C. and **L.-R. Dung** (2011*b*). A subsample-based low-power image compressor for capsule gastrointestinal endoscopy. *EURASIP Journal on Advances in Signal Processing*, **2011**(1), 257095.

Lin, M.-C., **L.-R. Dung**, and **P.-K. Weng** (2006). An ultra-low-power image compressor for capsule endoscope. *BioMedical Engineering OnLine*, **5**(1), 14.

Liu, G., **G. Yan**, **B. Zhu**, and **L. Lu** (2016). Design of a video capsule endoscopy system with low-power ASIC for monitoring gastrointestinal tract. *Medical & biological engineering & computing*, **54**(11), 1779–1791.

Liu, H., **N. Pan**, **H. Lu**, **E. Song**, **Q. Wang**, and **C.-C. Hung** (2013). Wireless

capsule endoscopy video reduction based on camera motion estimation. *Journal of digital imaging*, **26**(2), 287–301.

Liu, L., **S. Towfighian**, and **A. Hila** (2015). A review of locomotion systems for capsule endoscopy. *IEEE reviews in biomedical engineering*, **8**, 138–151.

Liu, X., **L. Wan**, **Y. Qu**, **T.-T. Wong**, **S. Lin**, **C.-S. Leung**, and **P.-A. Heng**, Intrinsic colorization. *In ACM SIGGRAPH Asia 2008 papers*. 2008, 1–9.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, **60**(2), 91–110.

Malathkar, N. V. and **S. K. Soni** (2019). Low complexity image compression algorithm based on hybrid DPCM for wireless capsule endoscopy. *Biomedical Signal Processing and Control*, **48**, 197–204.

Masci, J., **U. Meier**, **D. Cireşan**, and **J. Schmidhuber**, Stacked convolutional auto-encoders for hierarchical feature extraction. *In International Conference on Artificial Neural Networks*. Springer, 2011.

Mehmood, I., **M. Sajjad**, and **S. W. Baik** (2014). Video summarization based tele-endoscopy: a service to efficiently manage visual data generated during wireless capsule endoscopy procedure. *Journal of medical systems*, **38**(9), 1–9.

Memon, N., Adaptive coding of DCT coefficients by Golomb-Rice codes. *In Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, volume 1. IEEE, 1998.

Oliveira, P. A., **R. j. Cintra**, **F. M. bayer**, **S. Kulasekera**, and **A. Madanayake** (2017). Low-complexity image and video coding based on an approximate discrete tchebichef transform. *IEEE Transactions on Circuits and Systems for Video Technology*, **27**(5), 1066–1076.

Ou, G., **N. Shahidi**, **C. Galorport**, **O. Takach**, **T. Lee**, and **R. Enns** (2015). Effect of longer battery life on small bowel capsule endoscopy. *World Journal of Gastroenterology: WJG*, **21**(9), 2677.

Paschos, G. (2001). Perceptually uniform color spaces for color texture analysis: an empirical evaluation. *IEEE transactions on Image Processing*, **10**(6), 932–937.

Pennebaker, W. B. and **J. L. Mitchell**, *JPEG: Still image data compression standard*. Springer Science & Business Media, 1992.

Prattipati, S., **S. Ishwar**, **M. Swamy**, and **P. K. Meher**, A fast 8× 8 integer Tchebichef transform and comparison with integer cosine transform for image compression. *In 2013 IEEE 56th international midwest symposium on circuits and systems (MWSCAS)*. IEEE, 2013.

Primus, M. J., **K. Schoeffmann**, and **L. Böszörmenyi**, Segmentation of recorded endoscopic videos by detecting significant motion changes. *In 2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2013.

Rice, R. F. (1979). Some practical universal noiseless coding techniques.

Sargent, D., **C.-I. Chen**, **C.-M. Tsai**, **Y.-F. Wang**, and **D. Koppel**, Feature detector and descriptor for medical images. *In Medical Imaging 2009: Image Processing*, volume 7259. International Society for Optics and Photonics, 2009.

Shigemori, T. and **A. Matsui** (2011). Body-cavity image observation apparatus. US Patent 8,038,608.

Slepian, D. and **J. Wolf** (1973). Noiseless coding of correlated information sources. *IEEE Transactions on information Theory*, **19**(4), 471–480.

Smeaton, A. F., **P. Over**, and **A. R. Doherty** (2010). Video shot boundary detection: Seven years of TRECVid activity. *Computer Vision and Image Understanding*, **114**(4), 411–418.

Sullivan, G. J., **J.-R. Ohm**, **W.-J. Han**, and **T. Wiegand** (2012). Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on circuits and systems for video technology*, **22**(12), 1649–1668.

Thoné, J., **J. Verlinden**, and **R. Puers** (2010). An efficient hardware-optimized compression algorithm for wireless capsule endoscopy image transmission. *Procedia Engineering*, **5**, 208–211.

Tsevas, S., **D. K. Iakovidis**, **D. Maroulis**, and **E. Pavlakis**, Automatic frame reduction of wireless capsule endoscopy video. *In 2008 8th IEEE International Conference on BioInformatics and BioEngineering*. IEEE, 2008.

Turcza, P. and **M. Duplaga** (2011). Low power FPGA-based image processing core for wireless capsule endoscopy. *Sensors and Actuators A: Physical*, **172**(2), 552–560.

Turcza, P. and **M. Duplaga** (2013). Hardware-efficient low-power image processing system for wireless capsule endoscopy. *IEEE journal of biomedical and health informatics*, **17**(6), 1046–1056.

Turcza, P. and **M. Duplaga** (2017). Near-lossless energy-efficient image compression algorithm for wireless capsule endoscopy. *Biomedical Signal Processing and Control*, **38**, 1–8.

Varodayan, D., **Y.-C. Lin**, and **B. Girod** (2011). Adaptive distributed source coding. *IEEE transactions on image processing*, **21**(5), 2630–2640.

Wahid, K., **S.-B. Ko**, and **D. Teng**, Efficient hardware implementation of an image compressor for wireless capsule endoscopy applications. *In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. IEEE, 2008.

Wallace, G. K. (1992). The JPEG still picture compression standard. *IEEE transactions on consumer electronics*, **38**(1), xviii–xxxiv.

Wang, A., **S. Banerjee**, **B. A. Barth**, **Y. M. Bhat**, **S. Chauhan**, **K. T. Gottlieb**, **V. Konda**, **J. T. Maple**, **F. Murad**, **P. R. Pfau**, *et al.* (2013). Wireless capsule endoscopy. *Gastrointestinal endoscopy*, **78**(6), 805–815.

Welsh, T., **M. Ashikhmin**, and **K. Mueller**, Transferring color to greyscale images. *In Proceedings of the 29th annual conference on Computer graphics and interactive techniques*. 2002.

Wiegand, T., **G. J. Sullivan**, **G. Bjontegaard**, and **A. Luthra** (2003). Overview of the H. 264/AVC video coding standard. *IEEE Transactions on circuits and systems for video technology*, **13**(7), 560–576.

Wyner, A. and **J. Ziv** (1976). The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on information Theory*, **22**(1), 1–10.

Yatziv, L. and **G. Sapiro** (2006). Fast image and video colorization using chrominance blending. *IEEE transactions on image processing*, **15**(5), 1120–1129.

Yu, H., **Z. Lin**, and **F. Pan** (2005). Applications and improvement of H.264 in medical video compression. *IEEE Transactions on Circuits and Systems I: Regular Papers*, **52**(12), 2707–2716.

Yuan, Y. and **M. Q.-H. Meng**, Hierarchical key frames extraction for WCE video. *In 2013 IEEE International Conference on Mechatronics and Automation*. IEEE, 2013.

Yuan, Y., **J. Wang**, **B. Li**, and **M. Q.-H. Meng** (2015). Saliency based ulcer detection for wireless capsule endoscopy diagnosis. *IEEE transactions on medical imaging*, **34**(10), 2046–2057.

Zhang, R., **J.-Y. Zhu**, **P. Isola**, **X. Geng**, **A. S. Lin**, **T. Yu**, and **A. A. Efros** (2017). Real-time user-guided image colorization with learned deep priors. *arXiv preprint arXiv:1705.02999*.

Zhao, Q. and **M. Q.-H. Meng**, A strategy to abstract WCE video clips based on LDA. *In 2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011.

Zheng, Y., **L. Hawkins**, **J. Wolff**, **O. Goloubeva**, and **E. Goldberg** (2012). Detection of lesions during capsule endoscopy: Physician performance is disappointing. *Official journal of the American College of Gastroenterology— ACG*, **107**(4), 554–560.

Zhu, X., **A. K. Elmagarmid**, **X. Xue**, **L. Wu**, and **A. C. Catlin** (2005). Insightvideo: toward hierarchical video content organization for efficient browsing, summarization and retrieval. *IEEE Transactions on Multimedia*, **7**(4), 648–666.

Zhuang, Y., **Y. Rui**, **T. S. Huang**, and **S. Mehrotra**, Adaptive key frame extraction using unsupervised clustering. *In Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, volume 1. IEEE, 1998.

Zinaty, O., **E. Horn**, and **I. Bettesh** (2015). In-vivo imaging device providing data compression. US Patent 9,113,846.

List of Publications

Journal Publications

1. Sushma B and Aparna P. **Distributed video coding based on classification of frequency bands with block texture conditioned key frame encoder for wireless capsule endoscopy**, Biomedical Signal Processing and Control 60 (2020): 101940. (Status:Online)

2. Sushma B and Aparna P. **Summarization of wireless capsule endoscopy video using deep feature matching and motion analysis**, IEEE Access 9 (2020): 13691-13703. (Status: Online)

3. Sushma B and Aparna P. **Deep chroma prediction of WynerZiv frames in distributed video coding of wireless capsule endoscopy video**, Journal of Visual Communication and Image Representation 87 (2022): 103578. (Status: Online)

4. Sushma B and Aparna P. **Recent developments in wireless capsule endoscopy imaging: Compression and summarization techniques**, Computers in Biology and Medicine (2022): 106087. (Status: Online)

Conference Publications

1. Sushma B and Aparna P. **Texture Classification based Efficient Image Compression Algorithm for Wireless Capsule Endoscopy**, 5th International Conference ICCED 2019. (Status: Online)

Brief Bio-Data

Sushma B

Research Scholar.,

Department of Electronics and Communication Engineering

National Institute of Technology Karnataka, Surathkal

P.O. Srinivasnagar

Mangalore, 575025

Email: sushma.bg@gmail.com

Permanent Address

Sushma B

D/O Bhadra Reddy N

168, Kithiganahally

Bommasandra Industrial Area

Bangalore, 560099

Qualification

M. Tech. Digital Electronics and Advanced Communication, National Institute of Technology Karnataka, Surathkal, 2006.

B. E. Electronics and Communication Engineering, Siddaganga Institute of Technology, Tumkur, Karnataka, 2002.