AUTOMATIC CONTENT ANALYSIS OF ENDOSCOPY VIDEO (ENDOSCOPIC MULTIMEDIA INFORMATION SYSTEM)

by

SAE HWANG

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2007

To my adorable wife, Eunjoon, and my lovely children, Peter and Andy.

ACKNOWLEDGEMENTS

It is my pleasure to thank the many people who made this thesis possible.

I would like to gratefully acknowledge the enthusiastic supervision of Dr. Hua-mei Chen and Dr. JungHwan Oh to complete this thesis. My appreciation also goes to my other committee members: Dr. Sharma Chakravarthy, Dr. Roger Walker and Dr. Jean Gao. Thanks to all of them not just for being on my committee but for all of their guidance and assistance.

I would also like to thank all members in the Endoscopic Multimedia Information System (EMIS) who have helped me a lot throughout my Ph.D life. Especially, I want to say special 'Thank You' to Dr. JungHwan Oh, Dr. Wallapak Tavanapong, Dr. Piet C. de Groen, Dr. Johnny S. Wong and Yu Cao who have discussed my research topics whenever I encountered problems, and made persistent relationship with me.

I am forever indebted to my parents for providing a loving environment for me, and to my wonderful wife, Eunjoon Hwang for her sacrifice, encouragement and patience. She is my best friend in my life.

April 3, 2007

ABSTRACT

AUTOMATIC CONTENT ANALYSIS OF ENDOSCOPY VIDEO

(ENDOSCOPIC MULTIMEDIA INFORMATION SYSTEM)

Publication No. _____

Sae Hwang, Ph.D.

The University of Texas at Arlington, 2007

Supervising Professor: Hua-mei Chen & JungHwan Oh

Advances in video technology are being incorporated into today's healthcare practice. For example, various types of endoscopes are used for colonoscopy, upper gastrointestinal endoscopy, enteroscopy, bronchoscopy, cystoscopy, laparoscopy, and some minimal invasive surgeries (i.e., video endoscopic neurosurgery). These endoscopes come in various sizes, but all have a tiny video camera at the tip of the endoscopes. During an endoscopic procedure, the tiny video camera generates a video signal of the interior of the human organ, for example, the internal mucosa of the colon. The video data are displayed on a monitor for real-time analysis by the physician. Diagnosis, biopsy and therapeutic operations can be performed during the procedure. We define ***endoscopy videos*** as digital videos captured during endoscopic procedures.

Despite a large body of knowledge in medical image analysis, endoscopy videos are not systematically captured for real-time or post-procedure reviews and analyses. No hardware and software tools have been developed to capture, analyze, and provide user-friendly and efficient access to important content on such videos. To address this problem,

a project has been proposed to develop an ***Endoscopic Multimedia Information System (EMIS)*** which captures high quality endoscopy videos, analyzes the captured videos for important contents, and provides efficient access to these contents.

In this dissertation, we focus on the automatic analysis techniques of endoscopy videos for important contents by presenting (1) object & frame recognition, (2) multi-level endoscopy video segmentation and (3) application for endoscopy video analysis (Measurement of Endoscopy Quality). To analyze the contents of endoscopy videos, we first propose three object & frame recognition algorithm: *Endoscopy Video Frame Classification*, *Lumen Identification* and *Polyp Detection*.

The problem of segmenting visual data into smaller chunks is a basic problem in multimedia analysis, and its solution helps in problems such as video indexing and retrieval. However, traditional video segmentation techniques are not suitable for segmenting endoscopy video because endoscopy videos are generated by a single camera operation without shot, which makes it difficult to manage and analyze them. To address this problem, I propose a novel algorithm of multi-level segmentation for endoscopy video, which represents the semantic structure of endoscopy video: *Video*, *Phase*, *Piece*, and *Objective Shot*.

Based on the information obtained by object & frame recognition and multi-level endoscopy video segmentation, we develop software tool to measure the quality of endoscopic procedure. The development of software tool will directly benefit endoscopic research, education, and training: especially for the research-based advanced training of students in graduate and undergraduate programs in medical informatics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Advances in video technology are being incorporated into today's healthcare practices. Various types of endoscopes are used for colonoscopy, upper gastrointestinal endoscopy, enteroscopy, bronchoscopy, cystoscopy, laparoscopy, wireless capsule endoscopy, and some minimal invasive surgeries (i.e., video endoscopic neurosurgery). These endoscopes come in various sizes, but all have a tiny video camera at the tip of the endoscopes. During an endoscopic procedure, the tiny video camera generates a video signal of the interior of the human organ, which is displayed on a monitor for real-time analysis by the physician. We define ***endoscopy videos*** as digital videos captured during endoscopic procedures.

Colonoscopy is an important screening tool for colorectal cancer. In the US, colorectal cancer is the second leading cause of all cancer deaths behind lung cancer [1]. As the name implies, colorectal cancers are malignant tumors that develop in the colon and rectum. The survival rate is higher if the cancer is found and treated early before metastasis to lymph nodes or other organs occurs. The colon is a hollow, muscular tube about 6 feet long, as illustrated in Figure 1.1. A normal colon consists of six parts: cecum with appendix, ascending colon, transverse colon, descending colon, sigmoid and rectum. Colonoscopy allows for inspection of the entire colon and provides the ability to perform a number of therapeutic operations such as polyp removal during a single procedure. A colonoscopic procedure consists of two phases: *insertion phase* and *withdrawal phase*. During the insertion phase, a flexible endoscope (a flexible tube with a tiny video camera at the tip) is advanced under direct vision via the anus into the rectum and then grad-

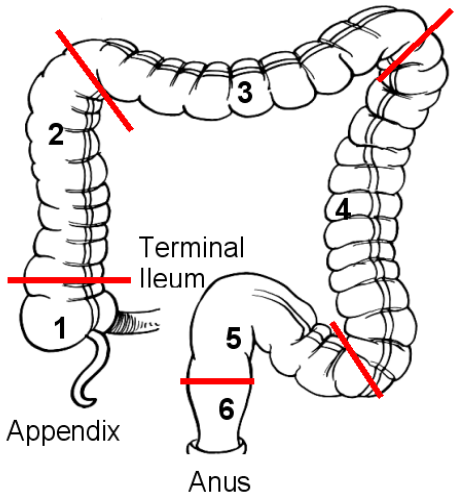

Figure 1.1. The colon endoscopic segments: 1-cecum, 2-ascending colon, 3-transverse colon, 4- descending colon, 5-sigmoid, 6-rectum clip.

ually into the most proximal part of the colon or the terminal ileum. In the withdrawal phase, the endoscope is gradually withdrawn [2, 3, 4]. The purpose of the insertion phase is to reach the end of colon (cecum or terminal ileum).Careful mucosa inspection and diagnostic or therapeutic interventions such as biopsy, polyp removal, etc., are performed in the withdrawal phase. Video data of colonoscopy are not routinely captured in the current practice. They typically have many out-of-focus frames. We call an out-of-focus frame a *non-informative frame*. The non-informative frames are usually generated due to two main reasons: too-close (or too-far) focus into (from) the mucosa of colon or foreign substances (i.e., stool, cleansing agent, air bubbles, etc.) covering camera lens or rapidly moving through the intracolonic space. This is because current endoscopes are equipped with a single, wide-angle lens that cannot be focused. Sharpness, brightness, and contrast of the video frames depend on the endoscopist's skills.

Colonoscopy is performed over 14 million times a year. Although colonoscopy has become the preferred screening modality for prevention of colorectal cancer, recent data suggest that there is a significant miss-rate for the detection of even large polyps and

cancers [5, 6, 7]. The miss-rate varies among endoscopists and it is known that the miss-rate is correlated with the level of the experience of an endoscopist defined as years performing the procedure. However, there is no measurement method to evaluate the endoscopist's skill and the quality of colonoscopic procedure. In general, the quality of a colonoscopic procedure can be evaluated in terms of the screening time of the withdrawal phase and the recognizability of a colonoscopy video of the withdrawal phase. Current American Society for Gastrointestinal Endoscopy (ASGE) guideline suggests that on average the withdrawal phase during a screening colonoscopy lasts a minimum of 6-10 minutes.

Despite the popularity of endoscopes and the promising evolution of image processing technology, there is very few research working on the novel solutions to analyze colonoscopy videos for important contents. In this dissertation, we focus on the automatic analysis techniques of endoscopy videos for important contents by presenting (1) three object & frame recognition (i.e. endoscopy video frame classification, lumen identification and polyp detection), (2) multi-level medical video segmentation and (3) application for medical video analysis (Measurement of Endoscopy Quality). The contributions of my dissertation can be presented as follows:

- Endoscopy Video Frame Classification: to distinguish non-informative frames from informative frames in endoscopy videos
  - Typically, a reference image is required to decide the quality (i.e., informative and non-informative) of an image. However, reference images are not available for a specific patient (each patient and each colon is unique). We propose two techniques that are able to evaluate the quality of an image without a reference image.
  - Since we do not use any domain knowledge of the video, the proposed technique is domain independent. Hence, it can be used for other medical videos

such as upper gastrointestinal endoscopy, enteroscopy, bronchoscopy, cystoscopy, and laparoscopy.

- Lumen Identification: to recognize lumen region in a frame and distinguish lumen-view frames from wall-view frames in endoscopy videos
  - The problem of deciding whether an image contains the distant colon lumen or not has not been investigated. A wall view occurs as a result of a close inspection of the colon wall whereas the lumen view indicates a more global inspection where more than one side of the colonic wall is within the field of vision. We propose a new lumen identification algorithm to decide whether an image has the colon lumen or not based on the bilateral convex shape of lumen
- Polyp Detection: to detect various sizes of polyps
  - We develop a new shape-based polyp detection method. Unlike the case in texture feature analysis, our method does not require system training which is very time consuming.
  - In articles dealing with CT Colography, general edge detection methods (Sobel edge detection, Canny edge detection, etc.) can be used because the boundary between colon wall and lumen is clear. However, they are not applicable for colonoscopy video images since the images have very complicated edges. Instead of edge detection, we use a marker-controlled watershed region segmentation which provides relatively clear edge information.
  - We propose new techniques to distinguish ellipses of polyp regions from those of non-polyp regions by matching curve direction, curvature, edge distance and intensity.
  - We develop a new method to segment colonoscopy videos into smaller parts by utilizing the mutual information based image registration technique. Each

part is a new type of semantic unit called *polyp shot*. The polyp shot detection method detects missed polyp frames by comparing the polyp candidate frames with their adjacent frames and determines the boundaries of polyp shots.

- Multi-level Medical Video Segmentation: to segment medical video representing the semantic structure of medical video using domain knowledge
  - Medical videos are usually generated by a single camera operation without shot. Thus, traditional video segmentation techniques cannot be applied to medical videos. To address this problem, we propose a new video segmentation technique to segment a colonoscopy video into phase, piece and object shot. Based on the analysis of camera motions, a colonoscopy video is segmented into insertion and withdrawal phases. Each phase is segmented into several pieces using our endoscopy video frame classification technique. Each piece can be decomposed into several kinds of shots based on human perception understanding the video contents such as endoscope movement and important objects. Objective shots are constructed by considering the spatio-temporal relationship within a video
- Application for Medical Video Analysis (Measurement of Endoscopy Quality): to produce objective measures of quality for colonoscopic procedures
  - We are the first to investigate an automatic measurement method that generates a number of objective metrics to evaluate the endoscopist's skill and the quality of colonoscopic procedures. Our metrics are developed based on expertise of a domain expert. No prior research has investigated objective quality measurement methods for colonoscopy or other endoscopic procedures.

The remainder of this paper is organized as follows. In Chapter 2, we present the background of this project. In Chapter 3, two techniques for endoscopy video frame classification (Edge-based and Clustering-based), are introduced in Section 3.1 and Sec-

tion 3.2, respectively. We discuss our experimental results in Section 3.3. In Chapter 4, we present the technique of the lumen identification. The lumen image identification and the lumen property determination are discussed in Section 4.1 and 4.2, respectively. We discuss our experimental results in Section 4.3. In Chapter 5, we present the polyp detection technique. The gradient magnitude construction is discussed in Section 5.1. The region segmentation based on the marker-controlled watershed algorithm in Section 5.2 and the polyp candidate selection is discussed in Section 5.3. In Section 5.4, we present the technique that generates polyp shots. We discuss our experimental results in Section 5.5. In Chapter 6, we present the technique of the multi-level endoscopy video segmentation. The camera motion estimation technique is presented in Section 6.1. The phase and motion shot segmentation technique based on camera motion estimation is presented in Section 6.2. We discuss our experimental results in Section 6.3. In Chapter 7, we present the technique of the measurement of endoscopy quality. In Section 7.1, we present the formal definitions of the quality metrics. We discuss our experimental results in Section 7.2.

CHAPTER 2

BACKGROUND

In this chapter, we discuss some previous works related with endoscopy and we present an overview of a project, ***Endoscopic Multimedia Information System (EMIS)***, to capture high quality endoscopy videos, analyze the captured videos for important contents, and provide efficient access to these contents.

2.1 Related Works

Despite intensive research in medical imaging in recent years, research on image analysis for colonoscopy videos has been minimal. Microrobotic endoscopy [8, 9, 10] focuses on identifying lumen boundary given that the image is known to have the lumen. Khan [8] proposed to use an N-level quadtree-based pyramid structure to find the most homogenous large dark region. Kumar *et al.* [9] proposed a global thresholding technique and differential region-growing to segment the lumen region. Tian et al. proposed APT-Iris that utilizes the relative darkness of the lumen [10]. Analysis of microscopic images from biopsies of tissues are described in [11, 12, 13]. However, microscopic images are different from endoscopic image. Computer-based approaches for the discrimination of gastric polyps have been proposed based on texture-feature. Color Wavelet Covariance generating a set of 72 texture features was proposed in [14]. The polyp detection techniques using the texture spectrum and Neural Network Classifier was proposed in [15]. More recently, the effectiveness of four different texture feature methods such as Texture Spectrum, Texture Spectrum with Color Histogram, Local Binary Pattern and Color Wavelet Covariance for detecting polyps was compared in [16]. Polyp detection in

CT colonography has received tremendous attention [17, 18, 19, 20]. CT colonography (CTC) or virtual colonoscopy is an emerging technology for acquisition and viewing of CT data sets created from an air-distended colon with helical CT scanners. CTC is a promising modality for screening for colorectal cancer. CTC is non-invasive and is performed without sedation; however, it also has disadvantages, making it not ready for practical screening. Poor bowl preparation or excess fluid remain in the bowl can result in false positives (retained stools are thought as polyps). Accuracy for polyp detection reported in the literature is varied significantly and there are safety concerns regarding to the cumulative exposure of patients to radiation with repeated surveillance [21]. The detection of flat and small polyps is still problematic. Nevertheless, the success in CTC is expected to drive up the demand for colonoscopy since polyps identified with CTC will require a subsequent colonoscopic procedure to remove them.

2.2 Endoscopic Multimedia Information System

Even though endoscopy has become very popular throughout the world, no hardware and software tools have been developed to capture, analyze, and provide user-friendly and efficient access to important content on such videos. To address this problem, a project has been proposed to develop an ***Endoscopic Multimedia Information System (EMIS)*** which captures high quality endoscopy videos, analyzes the captured videos for important contents, and provides efficient access to these contents. The *EMIS* team consists of University of Texas at Arlington (UTA), University of North Texas (UNT), Iowa State University (ISU), and Mayo Clinic Rochester (Mayo).

The development of software tool will directly benefit endoscopic research, education, and training: especially for the research-based advanced training of students in graduate and undergraduate programs in medical informatics.

CHAPTER 3

ENDOSCOPY VIDEO FRAME CLASSIFICATION

There are a significant number of out-of-focus frames in colonoscopy videos since current endoscopes are equipped with a single, wide-angle lens that cannot be focused. We define an ***out-of-focus*** frame as a ***non-informative frame*** (Figure 3.1) and an ***in-focus*** frame as an ***informative frame*** (Figure 3.2). The non-informative frames are usually generated due to two main reasons: too-close (or too-far) focus into (from) the mucosa of colon (the first two images in Figure 3.1) or foreign substances (i.e., stool, cleansing agent, air bubbles, etc.) covering camera lens or rapidly moving through the intracolonic space (the last two images in Figure 3.1). We call the procedure that distinguishes non-informative frames from informative frames in endoscopy videos ***Endoscopy Video Frame Classification*** in this paper. We propose two new techniques to distinguish non-informative frames from informative frames based on the detected edges, and Discrete Fourier Transform (DFT) with clustering, respectively. The edge-based approach is relatively simple and easy to implement, but sensitive to the selected threshold values. The DFT with clustering approach addresses the drawbacks of the edge-based approach, and provides more robust and accurate results.

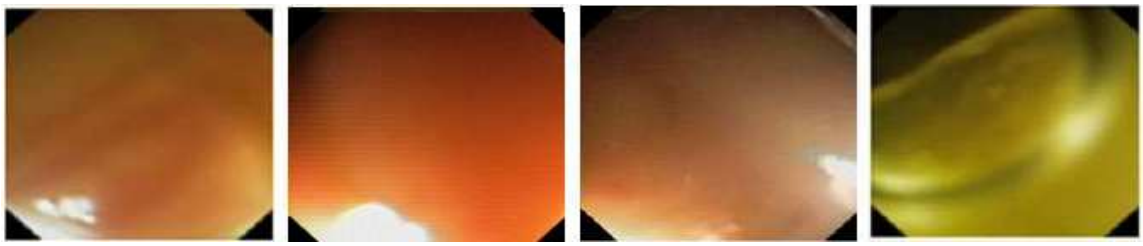

Figure 3.1. Examples of Non-Informative Frames.

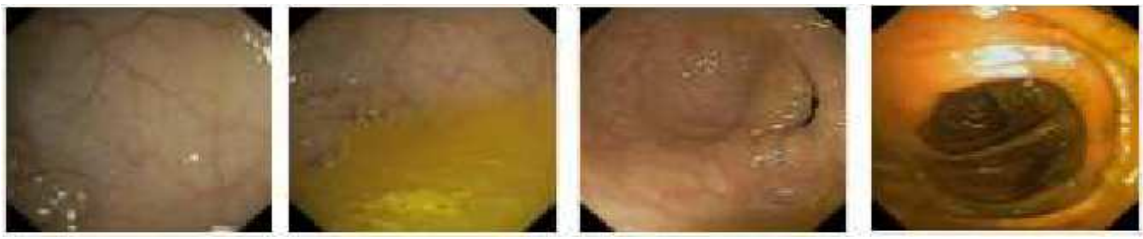

Figure 3.2. Examples of Informative Frames.

The output of endoscopy video frame classification provides information (i.e., frames that are informative) that will be used for further automatic or semi-automatic computer-aided diagnosis (CAD). It can reduce the number of images to be viewed by a physician and to be analyzed by a CAD system.

3.1 Edge-based Frame Classification

There are existing techniques [22, 23, 24, 25, 26, 27, 28] to handle out-of-focus images using image restoration. However, these existing techniques are not applicable to endoscopy video frames because these techniques need a reference image to compute the quality of the test image, and as already stated we only have test images. In this section, we propose a technique to distinguish non-informative frames from informative ones based on a property of isolated edge pixels.

We detect the edges from each frame using Canny Edge Detector [29]. Canny Edge Detector first smoothes an image to eliminate noise based on the Gaussian model and then tracks along the local maxima of the gradient magnitudes (edge strengths) of an image and sets to zero all pixels that are not actually the local maxima, which is known as non-maximal suppression. These two processes generate a single thin line for each edge when an image contains clear edge information and generate many isolated pixels when an image does not contain any clear edge information. Examples of the edge detection

results are shown in Figure 3.3, in which Figure 3.3 (b) and (c) are the images generated from applying the Canny Edge Detector on the image in Figure 3.3 (a). Figure 3.3 (f) and (g) show images generated from the image in Figure 3.3 (e). The parameters for the edge detector to generate images (b) and (f) are the same, but different from those used to generate images (c) and (g). As shown in this figure, the edge lines of the non-informative images are blurry and those of the informative images are clear regardless of the parameters used. The blurry lines occur due to discontinuity of the edge pixels constituting a line as seen in Figures 3.3 (d) and (h). Hence, to distinguish the blurry lines from the clear ones, we defined two terms, *isolated pixel* (IP) and *isolated pixel ratio* (IPR), for a frame as follows. An IP is an isolated edge pixel (edge pixel that is not connected to any other edge pixels) in a frame. We computed IPR as the percentage of the number of isolated edge pixels to the total number of edge pixels in the frame.

$$IPR = \frac{\text{Number of isolated pixels } (IPs)}{\text{Total number of pixels}} \times 100(\%) \tag{3.1}$$

The frame with the value of IPR greater than a certain threshold is declared a non-informative frame. Otherwise, the frame is considered an informative frame. However, there are some ambiguous images that can be either informative or non-informative according to the threshold value as seen in Figure 3.4 (a). This is because some images may have some parts that are blurry and other parts that are clear. For instance, in a tangential view along the mucosa, only some parts of the image will be clear. To handle these ambiguous images and optimize overall non-informative frame detection accuracy, we propose a two-step approach.

- Step 1: We classify frames into three categories: informative frames, non-informative frames and ambiguous frames using two very obvious thresholds for IPR, which are called the upper-threshold (TH_U) and the lower-threshold (TH_L). In other words,

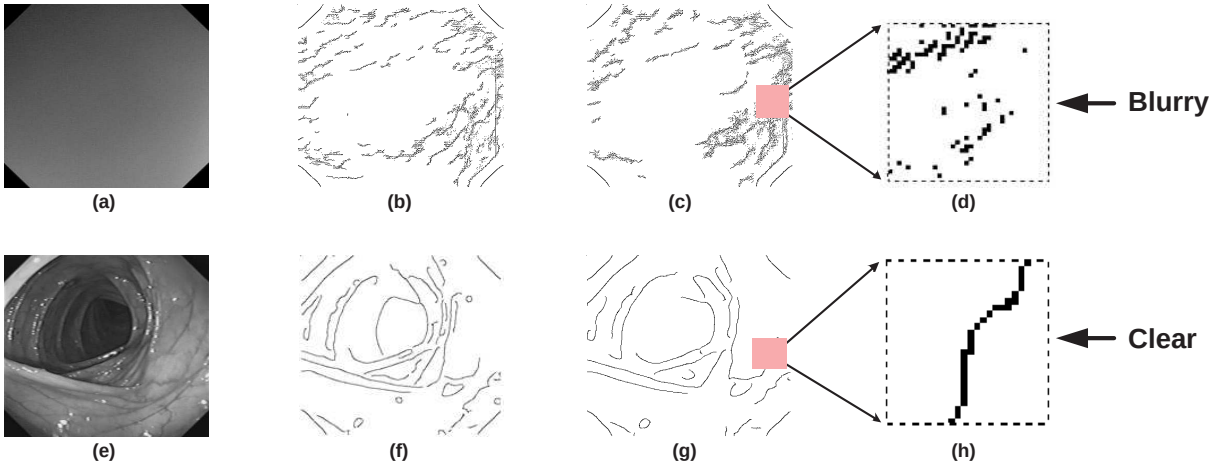

Figure 3.3. (a) Non-informative Image, (b) and (c) Edges detected from (a), (d) Details of Blurry Edge, (e) Informative Image, (f) and (g) Edges detected from (e), (h) Details of Clear Edge.

if an IPR of an image is larger than the upper-threshold value (TH_U), the image is classified as non-informative. If an IPR of an image is smaller than the lower-threshold value (TH_L), the image is classified as informative. If an IPR of an image is between upper-threshold and lower-threshold, the image is classified as ambiguous, and we proceed to Step 2.

- Step 2: An ambiguous frame is divided into a number (64 in our case) of blocks as seen in Figure 3.4 (b). First, each block is classified as empty or non-empty block. An empty block has no pixels. A non-empty block is classified into a clear or blurry block again. For block classification, we use only the lower-threshold value. If a frame has more informative blocks than non-informative ones, then it is classified as an informative frame.

Figure 3.4. (a) Ambiguous Frame, (b) Edge detected from (a) with 64 Blocks.

3.2 Clustering-based Frame Classification

The edge-based informative frame classification algorithm shows good performance results. However, there is a major drawback in this approach, which is that the performance of our edge-based technique is susceptible to the appropriate values of various parameters (i.e., sigma, high, low, etc.) in the edge detection algorithm, and the upper and lower thresholds in Step 1 and Step 2 of Section 3.1. To address this, we investigate a new approach based on Discrete Fourier Transform (DFT), texture analysis and data clustering. Figure 3.5 shows the framework of the proposed algorithm.

3.2.1 Feature Extraction

The basic idea used to detect informative frames comes from Discrete Fourier Transform (DFT) and texture analysis of its frequency spectrum. The process of DFT for a 2D image is that first, an image such as Figure 3.6 (a) or Figure 3.7 (a) is converted into the grayscale image, and then the grayscale image is transformed using the Fourier Transform [30, 31, 32, 33, 34]. The frequency spectrum, 2D plot of the magnitude of the Fourier Transform, is constructed using the coefficients of the Fourier Transform of a grayscale image. The frequency spectrum shows the frequency distribution of an image (Figure 3.6

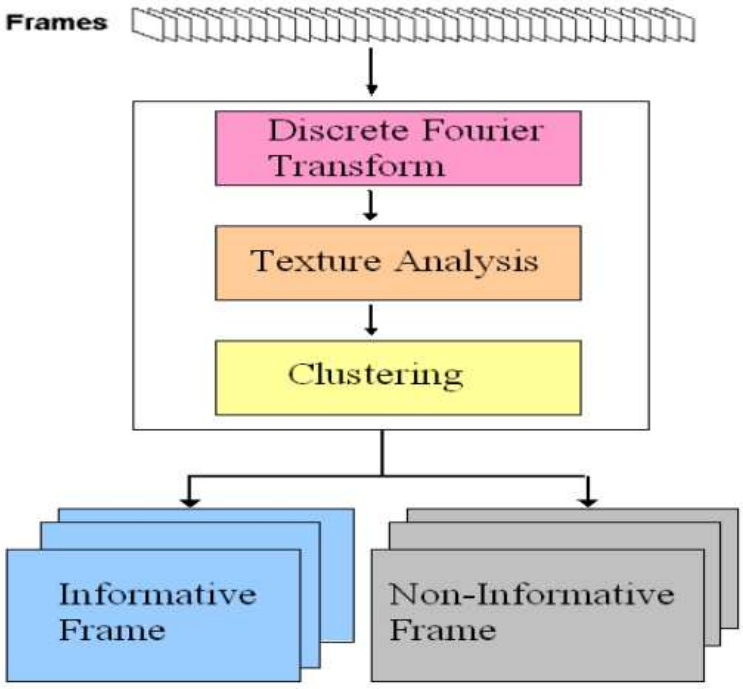

Figure 3.5. Framework of Informative and Non-Informative Frame Classification.

(b) or Figure 3.7 (b)). Based on the contents of the image, the frequency spectrums generate different patterns. It is usually impossible to make direct associations between specific components of an image and its transform. However, some general statements can be made about the relationship between the frequency components of the Fourier transform and spatial characteristics of an image. Typically, high frequencies hold the information of fluctuations of edges and boundaries, and low frequencies correspond to the slowly varying components of an image. The non-informative frame (Figure 3.6 (a)) has no clear object information except the 4 strong edges at the corners of an image running approximately at $\pm 45^{\circ}$ so its Fourier spectrum (Figure 3.6 (b)) shows prominent components along the $\pm 45^{\circ}$ directions that correspond to the 4 corners of an image. Compared to the non-informative frame, the informative frame (Figure 3.7 (a)) has a lot of clear edge information so its spectrum (Figure 3.7 (b)) of the informative frame does

not show prominent components along the ±45° directions because it has a wider range of bandwidth from low to high frequencies.

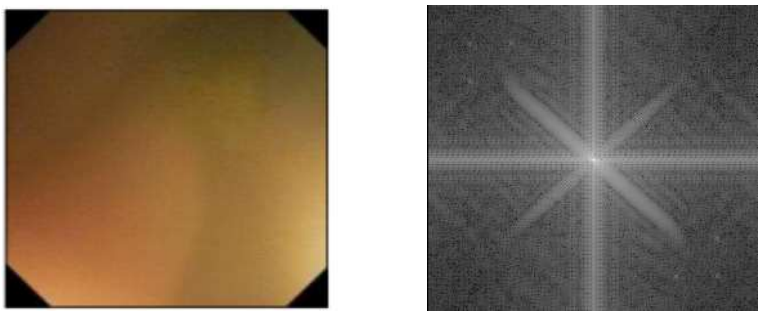

Figure 3.6. (a) Non-Informative Frame, (b) Frequency Spectrum of (a).

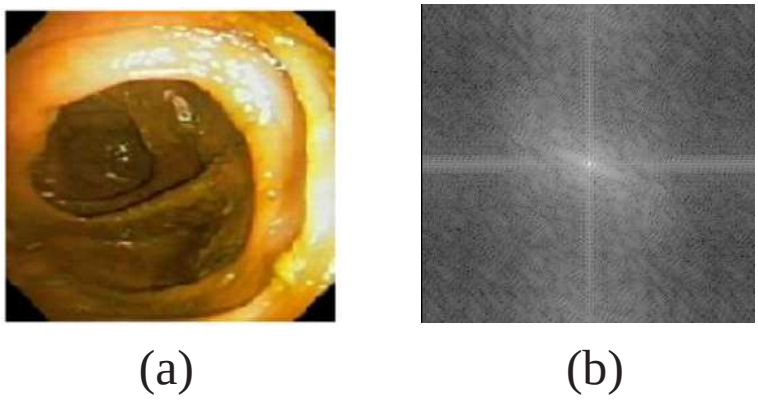

Figure 3.7. (a) Informative Frame, (b) Frequency Spectrum of (a).

3.2.2 Texture Analysis

The texture analysis is applied on the frequency spectrum image, which is a 2D plot of the magnitude, in order to find the pattern difference between the informative and the non-informative frames. The most well known statistical approach toward texture

analysis is the gray level co-occurrence matrix (GLCM) [35, 12, 36, 37, 38, 39, 40]. The co-occurrence matrix contains the elements that are the counts of the number of pixel pairs for specific brightness levels, when separated by some distance (or displacement) at some relative inclination. To construct the co-occurrence matrix for this texture analysis, we set up a window (matrix) of size equal to the size of the frequency spectrum image itself, a displacement to 1, and a relative inclination to 0. The original investigation into the texture features based on the co-occurrence matrix was pioneered by Haralick et al [35]. They defined 14 texture features. However, only some features among 14 texture features are in wide use in many applications [38, 39]. For our experiments, seven texture features (Entropy, Contrast, Correlation, Homogeneity, Dissimilarity, Angular Second Moment, and Energy) are extracted as follows [40].

$$\text{Entropy: } \sum_i \sum_j P(i,j) \cdot \log P(i,j) \tag{3.2}$$

$$\text{Contrast: } \sum_i \sum_j (i,j)^2 \cdot P(i,j) \tag{3.3}$$

$$\text{Correlation: } \sum_i \sum_j \frac{(i-\mu_x) \cdot (j-\mu_y) \cdot P(i,j)}{\sigma_x \sigma_y} \tag{3.4}$$

$$\text{Homogeneity: } \sum_i \sum_j \frac{P(i,j)}{1- \mid i-j \mid} \tag{3.5}$$

$$\text{Dissimilarity: } \sum_i \sum_j P(i,j) \cdot \mid i-j \mid \tag{3.6}$$

$$\text{Angular Second Moment (ASM): } \sum_i \sum_j P(i,j)^2 \tag{3.7}$$

$$\text{Energy: } \sqrt{ASM} \tag{3.8}$$

where $P(i,j)$ is the probability of a certain value in the co-occurrence matrix, $\mu_x = \sum_i \sum_j i \cdot P(i,j)$, $\mu_y = \sum_i \sum_j j \cdot P(i,j)$, $\sigma_x = \sqrt{\sum_i \sum_j (i-\mu_x)^2 \cdot P(i,j)}$ and $\sigma_y = \sqrt{\sum_i \sum_j (j-\mu_y)^2 \cdot P(i,j)}$. The extracted seven texture features are used to distinguish

the non-informative from the informative frames in the colonoscopy video using K-means clustering algorithm.

3.2.3 Clustering-based Classification

The K-means method is a well-known partitioning method, and is commonly used [41, 42, 43, 44, 45, 46]. The K-means method clusters data objects into K subsets using a certain distance function, where data objects in the same cluster are similar to one another but data objects in other clusters are dissimilar. Figure 3.8 describes the K-means clustering algorithm when a data object (X_i) consists of p dimensional features (i.e., $X_i = \{x_i^1, x_i^2, \cdots, x_i^p\}$).

K-Means Algorithm

---

1. Initialization - randomly choose K points $C_1, C_2, \cdots, C_k$ as initial centroids, where an initial centroid(C_j) is
   $C_j = \{c_j^1, c_j^2, \cdots, c_j^p\}$
2. Repeat
   (a) For i=1 to n (n = the total number of data objects)
   compute distance $d_j(X_i) = \sqrt{(x_i^1 - c_j^1)^2 + (x_i^2 - c_j^2)^2 + \cdots + (x_i^p - c_j^p)^2}$
   assign X_i to cluster D_{j*} where $j* = \min(d_1(X_i), d_2(X_i), \cdots, d_k(X_i))$
   (b) Compute the new centroid (C_j) for each cluster $D_j, j = 1, 2, \cdots, k$
   $C_j = \frac{1}{|D_j|} \sum_{X_i \in D_j} X_i = \left(\frac{\sum x_i^1}{|D_j|}, \frac{\sum x_i^2}{|D_j|}, \cdots, \frac{\sum x_i^p}{|D_j|} \right)$
   where $|D_j|$ is the number of data objects in cluster D_j
   (c) Exit, if the centroids no longer move

Figure 3.8. K-means Clustering Algorithm.

For our purpose, it is natural to set up the initial number of clusters to 2 ($k = 2$) and cluster the frames into two groups. One represents the informative frame group, and the other represents the non-informative frame group. We call this approach a one-step

K-means clustering scheme. Even though the one-step K-means clustering scheme distinguishes the informative frame from the non-informative frame very well, we investigate whether a larger number of initial clusters (k) can further increase its overall accuracy. There are frames in which some parts are clear, but other parts are blurry. As before, we call these frames *ambiguous* frames. Figures 3.9, 3.10 and 3.11 show three types of frames (Non-informative, Informative and Ambiguous).

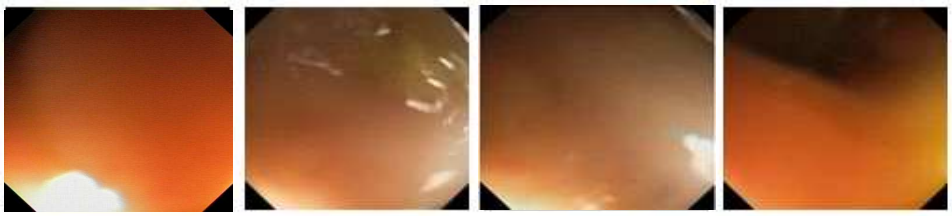

Figure 3.9. Examples of Non-Informative Frames.

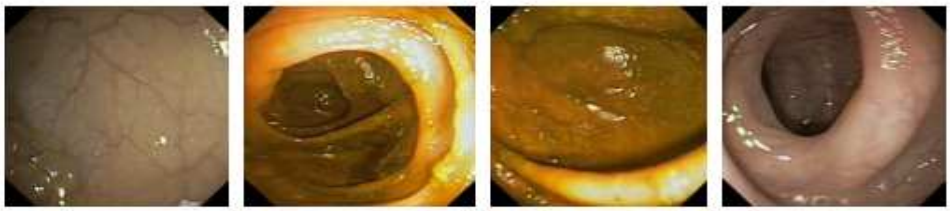

Figure 3.10. Examples of Informative Frames.

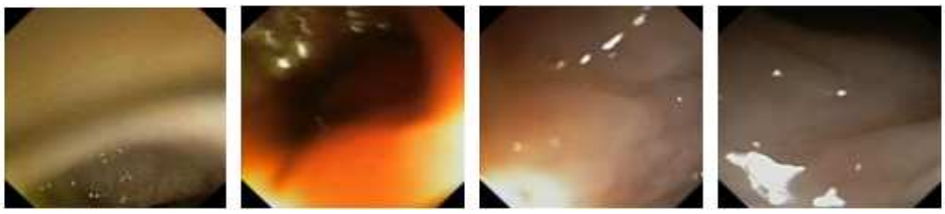

Figure 3.11. Examples of Ambiguous Frames.

Analogous to the edge-based method, we next develop a two-step K-means clustering scheme to distinguish the informative frames from non-informative frames. In the first clustering step, we set the initial number of clusters to 3 ($k = 3$) in order to cluster frames into three groups: informative frames, non-informative frames, and ambiguous frames. The frames detected as ambiguous from the first step are used in the next clustering step. In the second clustering step, we set up the number of clusters to 2 ($k = 2$) in order to further divide the ambiguous frames into two groups that consist of informative frames and non-informative frames. Finally all frames are clustered into two groups, either the informative frame or the non-informative frame groups. Our experiment results show that the two-step K-means clustering scheme is better than the one-step K-means clustering scheme.

3.3 Experimental Results

Our experiments assess the performances of the two proposed techniques for edge-based and clustering-based classification. To verify the effectiveness of our proposed algorithms, four traditional performance metrics [41] such as precision, sensitivity (recall), specificity, and accuracy, are measured in our experiments. Those four performance metrics are described as follows.

Table 3.1. Four Data Classification for Performance Metrics

	Predicted as Positive	Predicted as Negative
Actually Positive	TP	FN
Actually Negative	FP	TN

$$\text{Precision} = \frac{\text{Number of correct positive Predictions}}{\text{Number of Predictions}} = \frac{TP}{TP+FP}$$
$$\text{Recall} = \frac{\text{Number of Correct Positive Predictions}}{\text{Number of Positives}} = \frac{TP}{TP+FN}$$
$$\text{Specificity} = \frac{\text{Number of Correct Negative Predictions}}{\text{Number of Negative}} = \frac{TN}{FP+TN}$$
$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Number of Predictions}} = \frac{TP+TN}{TP+TN+FP+FN}$$

We note that the resolutions of the original images are 391 x 375 and 571 x 451. However, odd lines (or even lines) in both horizontal and vertical directions are removed, and the images are resized from 391 x 375 to 195 x 187 and 571 x 451 to 285 x 225 to reduce degradation by interlacing.

3.3.1 Evaluation of Edge-based Frame Classification

To distinguish informative frames from non-informative frames using the proposed edge-based method, we need to decide the upper-threshold (TH_U) and the lower-threshold (TH_L) values as mentioned in Section 3.1. We examined two sample data sets, each of which contains 2000 frames, to determine the thresholds. The size of the frames in the first set is 285x225 pixels and that of the second set is 195x187 pixels. Each frame of the data sets is classified into one of the three categories (informative frame, non-informative frame and ambiguous frame) manually based on the quality of the images. The results of this manual classification for the two sample data sets can be seen in Table 3.2 as follows.

The *IPR* value for each frame in the two data sets is computed. The Minimum, Maximum, Average and Median values of *IPR* for each category of the data sets are shown in Tables 3.3 and 3.4. For illustration purpose, the distribution of the *IPR* values of 2000 frames is presented in Figure 3.12 and 3.12. As seen in Tables 3.3 and 3.4, most of the informative frames have the low *IPR* values such that the average *IPR* of

Table 3.2. Manual Classification of Two Sample Data Sets

	Set 1 (285 x 225)	Set 2 (195 x 187)
# of informative frames	1479	1157
# of non-informative frames	258	646
# of ambiguous frames	263	197
Total	2000	2000

informative frames is around 1%, and the maximum *IPR* of informative frames is less than 5%. In contrast, the ambiguous frames and the non-informative frames have higher *IPR* values such that the average *IPR* of non-informative frames is around 6 to 7%, and the average *IPR* of ambiguous frames is around 3 to 5%.

Table 3.3. Statistics of Data Set 1 (285 x 225)

IPR (%)	Informative (*IPR*)	Non-informative (*IPR*)	Ambiguous (*IPR*)
Minimum	0.016	1.725	0.460
Maximum	4.926	10.451	9.155
Average	0.849	7.291	4.615
Median	0.541	7.455	4.387

Table 3.4. Statistics of Data Set 2 (195 x 187)

IPR (%)	Informative (*IPR*)	Non-informative (*IPR*)	Ambiguous (*IPR*)
Minimum	0.000	0.222	0.133
Maximum	4.930	12.130	7.821
Average	0.753	5.982	3.137
Median	0.401	6.538	3.000

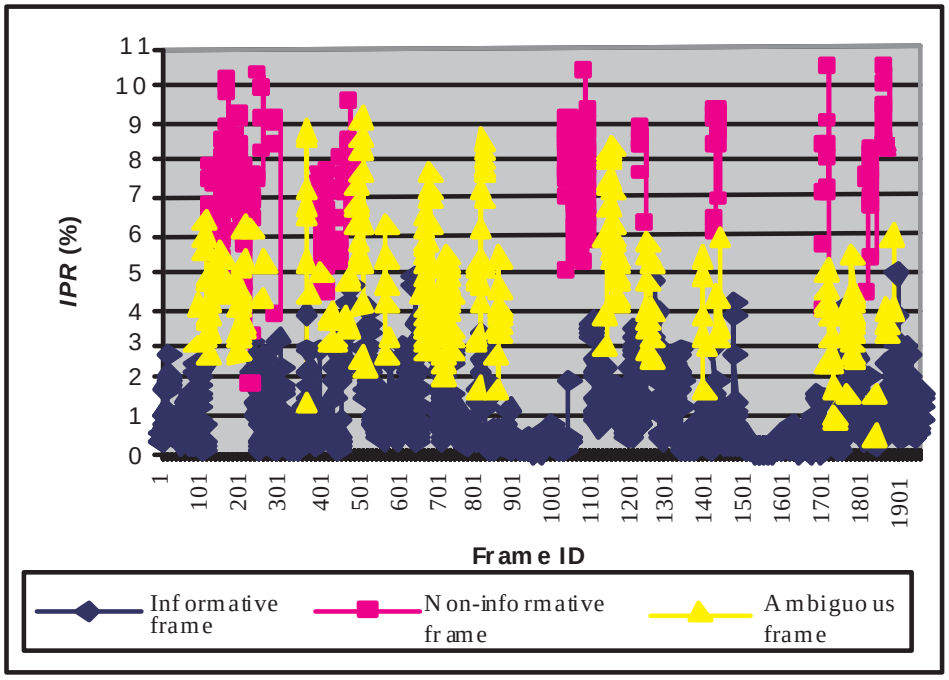

Figure 3.12. Distribution of *IPR* of Data Set 1.

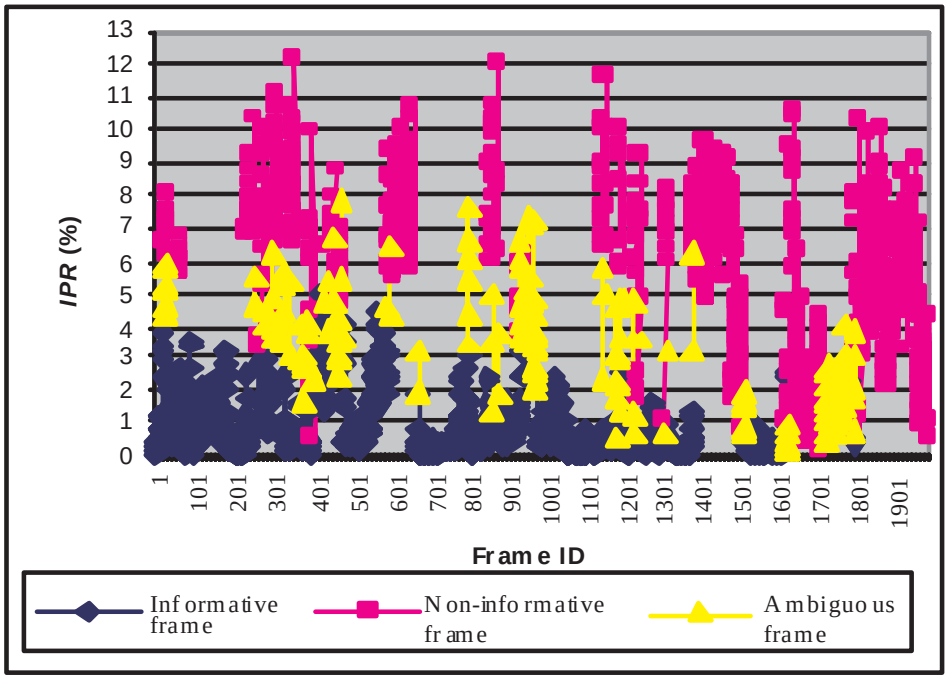

Figure 3.13. Distribution of *IPR* of Data Set 2.

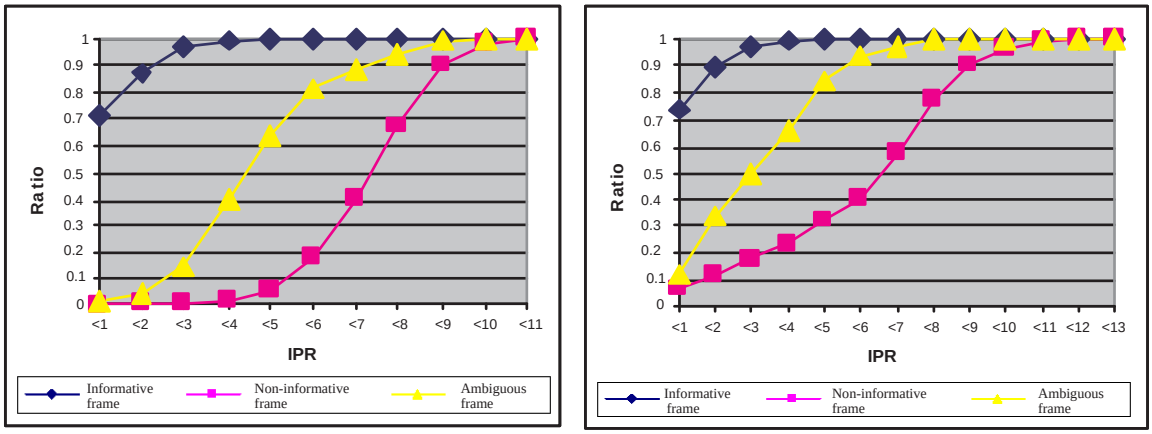

Figure 3.14. Accumulated Ratios of Informative Frames, Non-informative Frame and Ambiguous Frames for Data Set 1 (left) and Data Set 2 (right).

Figure 3.14 shows the accumulated ratios of the number of informative frames, non-informative frames and ambiguous frames of each data set based on *IPR* values. As shown in this figure, the *IPR* values of all informative frames are less than 5%. However, the *IPR* values of all non-informative and ambiguous frames are distributed over a wide range (from less than 1% to more than 12%). Therefore, we select the two threshold values as follows.

- The candidates for the lower-threshold (TH_L) value should be less than 5% because all informative frames have the *IPR* values less than 5%. The intuitive criterion for the TH_L is that the portion of detected informative frames by the selected TH_L should be greater than that of the detected non-informative and ambiguous frames. This comparison can be done by computing the difference between the ratio of the number of informative frames and the ratio of the number of non-informative and ambiguous frames. The difference (D^{IPR}) for an *IPR* value, i, is calculated as follows.

  $D_i^{IPR} = CR_i - (BR_i + AR_i)$

where CR_i is the ratio of the number of informative frames, BR_i is the ratio of the number of non-informative frames, and AR_i is the ratio of the number of ambiguous frames at IPR i. The subtraction works here since each value is a ratio which is not an absolute but a relative value. The results for the IPR 1, 2, 3, 4 and 5% are illustrated in Table 3.5. In our experiment, IPR 1, 2 and 3% are selected as TH_L values since the differences (D^{IPR}) of the three are much larger than those of the others.

Table 3.5. Results of Differential Calculation for Low Threshold

IPR	D^{IPR} of Set 1	D^{IPR} of Set 2	Average D^{IPR}
<1	0.699	0.554	**0.6265**
<2	0.831	0.442	**0.6365**
<3	0.825	0.299	**0.5620**
<4	0.576	0.097	**0.3365**
<5	0.314	-0.170	**0.0720**

- The candidates for the upper-threshold (TH_U) value should be selected greater than or equal to 5 % because all informative frames have the IPR values less than 5 %. Since we already determined the lower-threshold (TH_L) values as 1, 2, or 3 %, we ran experiments with different pairs of TH_U and TH_L values such as 5, 6, 7, and 8 for TH_U and 1, 2 and 3 for TH_L to determine the optimal TH_U value. The results are shown in Figure 3.15. As seen in the figure, there is little change in the number of frames detected as informative even if TH_U values are changing from 5 to 8. For example, in the first graph, about 1450 and 1330 frames are detected as informative frames when TH_L is 1 for the sample dataset 1 and 2 respectively, irrespective of TH_U values, which are ranged from 5 to 8. In the second graph, about 1600 and 1440 frames are detected as informative frames when TH_L is 2,

and about 1680 and 1520 frames are detected as informative frames when TH_L is 3 in the third graph for the sample dataset 1 and 2 respectively.

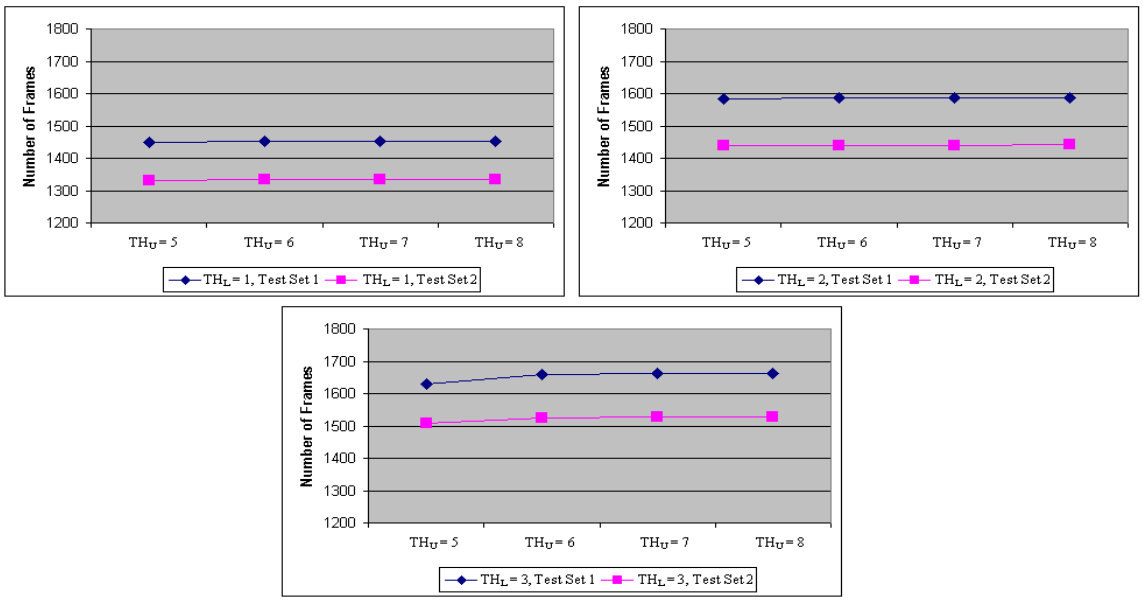

Figure 3.15. Detected Informative Frames based on Different Pairs of Thresholds.

Using a set of threshold values determined above (1, 2 and 3 for TH_L , and 5, 6, 7, and 8 for TH_U), we have run our edge-based non-informative frame detection algorithm. The overall results for the precision and the recall are summarized in Table 3.6 compared with several combinations of the low-threshold (TH_L) from 1 to 3 and the upper-threshold (TH_U) from 5 to 8. The 'Average' in Table 3.6 is an average value of the precision and sensitivity. As seen in the table, the results are very good, and the accuracy does not vary much with the threshold values.

We applied our edge-based technique to the two colonoscopy video test sets. The actual video frame rate of our colonoscopy video is 30 frames per second. However, we extracted frames at the rate of 1 frame per second because the evaluation is performed on individual frames so the extraction rate will not become a performance degrading factor.

Table 3.6. Precision and Sensitivity based on Several Combinations of Thresholds

Thresholds	Test Set 1			Test Set 2		
	Precision	Sensitivity	Average	Precision	Sensitivity	Average
TH_L=1, TH_U=5	1.000	0.936	0.968	0.916	0.965	0.940
TH_L=2, TH_U=5	0.979	1.000	0.989	0.898	0.996	0.947
TH_L=3, TH_U=5	0.949	1.000	0.974	0.869	1.000	0.934
TH_L=1, TH_U=6	1.000	0.936	0.968	0.915	0.965	0.940
TH_L=2, TH_U=6	0.976	1.000	0.988	0.897	0.996	0.946
TH_L=3, TH_U=6	0.934	1.000	0.967	0.859	1.000	0.929
TH_L=1, TH_U=7	1.000	0.936	0.968	0.915	0.965	0.940
TH_L=2, TH_U=7	0.976	1.000	0.988	0.897	0.996	0.946
TH_L=3, TH_U=7	0.932	1.000	0.966	0.857	1.000	0.928
TH_L=1, TH_U=8	1.000	0.936	0.968	0.915	0.965	0.940
TH_L=2, TH_U=8	0.975	1.000	0.987	0.897	0.996	0.947
TH_L=3, TH_U=8	0.930	1.000	0.966	0.856	1.000	0.928

The total length of videos in our test set is about 15 minutes and the test set consists of 923 frames. There are two different resolutions (285 x 225 and 195 x 187pixels) in our videos. The details about our test video set can be found in Table 3.7.

Table 3.7. Test Set of Videos

Video ID	Video Length (min)	Total # of Frames	Resolution
Colon-1	10	627	285 x 225
Colon-2	5	296	195 x 187
Total	15	923	

Figure 3.16 shows the experimental results of our edge-based non-informative frame classification technique. The results indicate the proposed technique is acceptable achieving over 88% for four different performance metrics (i.e. precision, sensitivity, specificity, and accuracy).

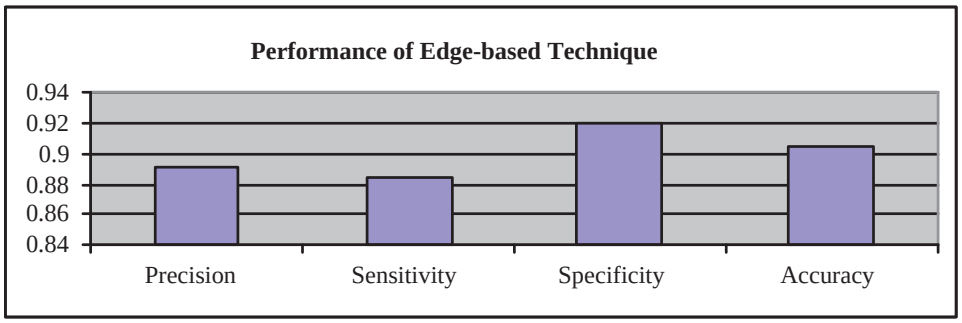

Figure 3.16. Performance of Edge-based Technique.

3.3.2 Evaluation of Clustering-based Frame Classification

Next, we studied the performance of each of the seven texture features and compared the performance of the one-step and the two-step clustering schemes. And, we examined how effectively the specular reflection detection technique could increase the performance of the clustering-based classification technique. The data set used in this section is the same test video (two colonoscopies) set described in Table 3.7 of Section 3.3.1. First, we examined the individual performance of each of the seven texture features to see if there is a dominant texture feature distinguishing non-informative frames from informative frames. We also present the performance of all seven features used together. Figure 3.17 shows each performance metric of the one-step clustering scheme and Figure 3.18 shows each performance metric of the two-step clustering scheme. The labels in the x-coordinate represent the name of texture features and the label of '7 Features' means that all seven features are used together. 'Colon-1' and 'Colon-2' in the legend indicate the video ID, and 'Ave' in the legend means the average performance metrics of two colonoscopy videos. Figure 3.17 and Figure 3.18 show that the performance of all seven features used together is better than performances of individual texture features for both the one-step and the two-step clustering schemes. We note that the two-step

clustering scheme provides better results than the one-step clustering scheme and that the combined use of all seven features optimizes the results.

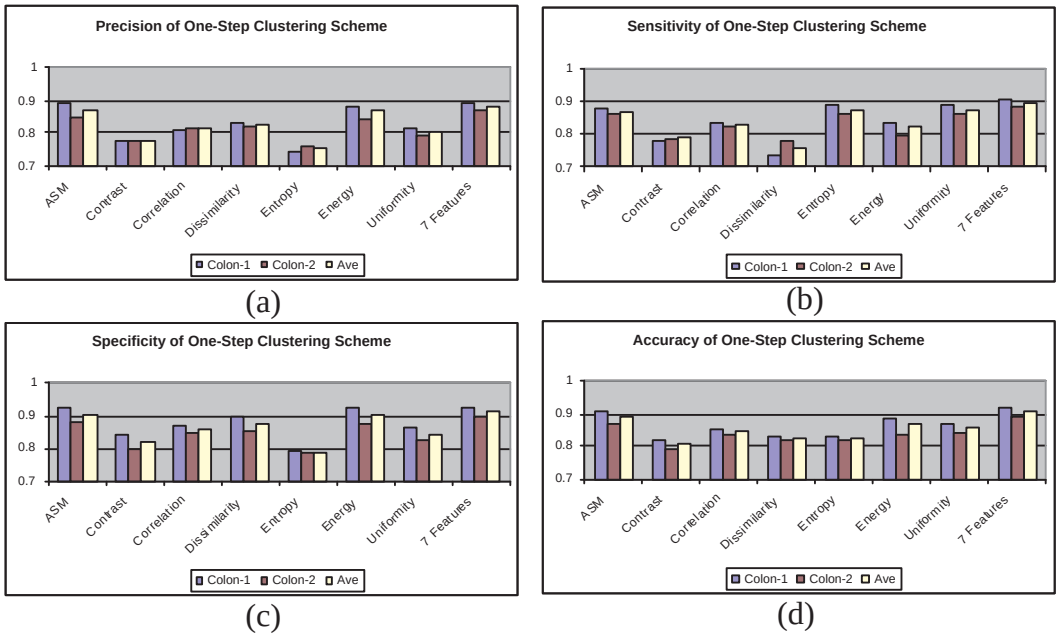

Figure 3.17. Effectiveness of Different Texture Features on Performance of One Step Clustering Scheme. (a) Precision, (b) Sensitivity, (c) Specificity, (d) Accuracy .

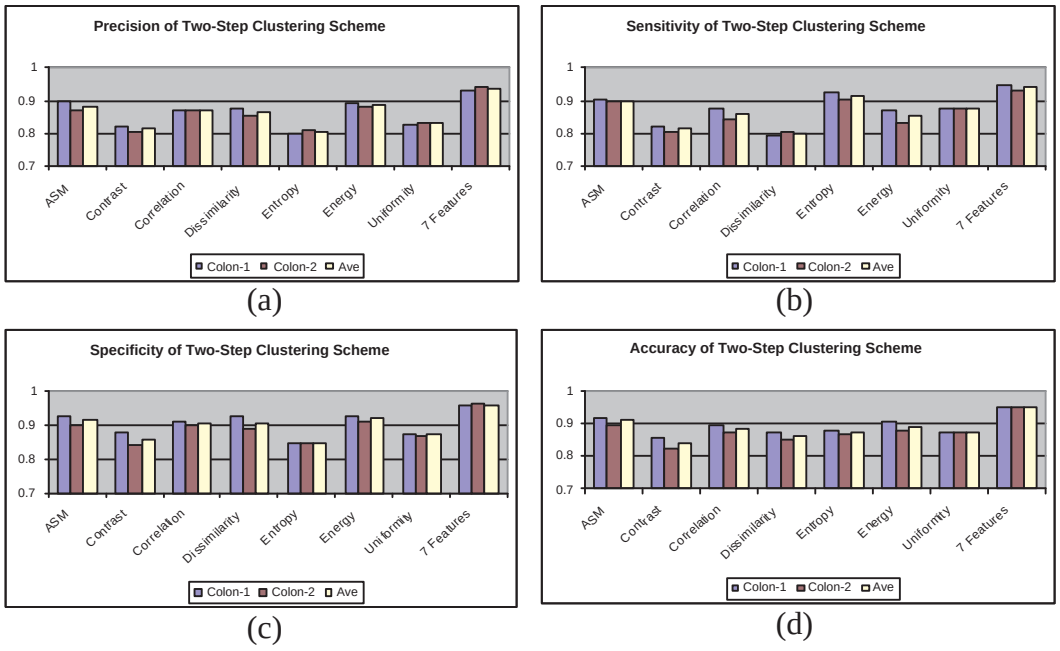

Figure 3.18. Effectiveness of Different Texture Features on Performance of Two Step Clustering Scheme. (a) Precision, (b) Sensitivity, (c) Specificity, (d) Accuracy .

CHAPTER 4

LUMEN IDENTIFICATION

Lumen identification is used to derive the metric to evaluate mucosa inspection during the withdrawal phase. A *lumen view* is defined as an informative frame that contains the colon lumen whereas an informative frame without the colon lumen is called a *wall view*. A wall view occurs as a result of a close inspection of the colon wall whereas the lumen view indicates a more global inspection where more than one side of the colonic wall is within the field of vision. The problem of deciding whether an image contains the distant colon lumen or not has not been investigated in the literature. The most related research effort determines the colon lumen boundary given an image with the colon lumen for microrobotic endoscopy in [10]. We propose lumen identification technique based on bilateral convex shape of lumen using the following steps (Figure 4.1). In step 1, we identify whether a frame is an informative frame or not using our endoscopy video frame classification technique. In step 2, for each informative frame, we identify whether the frame have the colon lumen or not. In step 3, we determine the lumen properties (size and location) for each frame with the colon lumen.

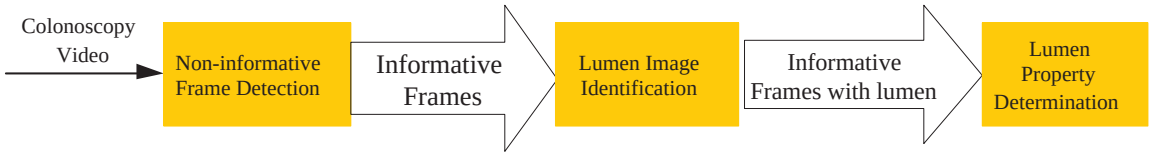

Figure 4.1. Framework of Lumen Identification.

4.1 Lumen Image Identification

After we filter out the non-informative frames, we identify whether the frame have the colon lumen or not. The colon lumen is relatively darker ($R1$ in Figure 4.2 (b)) and there is more than one bilateral convex colon wall around the colon lumen. The intensity difference between consecutive colon walls is small. We design our technique based on this observation. Our technique utilizes the algorithm in [47] to determine whether a planar region is convex or concave. Region R is considered convex if and only if for any pair of points p and q in R, the line segment connecting p and q, is completely in R; otherwise, the region is considered concave. Our technique works as follows.

1. Segment the image using JSEG [48] and filter out all the regions whose size is less than a pre-defined size threshold t_1. This is to eliminate regions that are too small and unlikely to be the distant colon lumen.
2. Let r_1 represent the region with the lowest pixel intensity initially. If the intensity of r_1 is greater than another intensity threshold t_2 or r_1 is concave, declare that this image is a wall view (no colon lumen). Otherwise, we check for two colon walls surrounding the colon lumen as follows.

   Step 1: Let r_2 be the closest neighboring concave region of r_1. Compare the intensity difference between r_1 and r_2. If the difference is larger than the intensity difference threshold t_3, declare that this image is a wall view and the algorithm terminates. Otherwise, proceed to Step 2.

   Step 2: Let r_1 denote the region r_2 and proceed to Step 1 if this is the first time Step 2 is executed. Otherwise, declare that the image is a lumen view and the algorithm terminates.

   Note that we repeat the two steps twice to check that at least two colon walls are seen together with the colon lumen before we declare that the image is a lumen view.

Figure 4.2 (a) shows an original image with the colon lumen almost in the center. Figure 4.2 (b) depicts the segmented image with the important regions, $R1, R2$, and $R3$ labeled. For ease of visualization, we generate Figure 4.2 (c) by masking small regions and neighboring convex regions with black pixels. Figure 4.2 (c) shows only $R1, R2$, and $R3$. Region $R1$ is the convex region with the smallest intensity, representing the distant colon lumen. $R2$ is the concave region close to $R1$, representing a segment of the colon wall. Considering $R1$ and $R2$ together, we see a bilateral convex colon wall. $R3$ is another concave region close to $R2$, representing another segment of the colon wall.

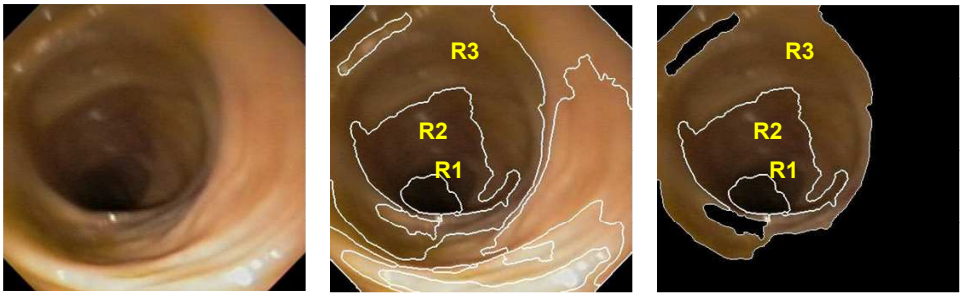

Figure 4.2. Images processed during lumen identification.

4.2 Lumen Property Determination

In this section, we determine the lumen properties (size and location) for each frame with the colon lumen. Fist, we select the initial lumen region using the Otsu's thresholding algorithm. After then, we obtained the elliptical lumen region using ellipse fitting method.

4.2.1 Initial Lumen Region Selection using Otsu Thresholding

We apply the Otsu's thresholding algorithm iteratively like in the APT-Iris technique [10]. We stop at the iteration when Equation (4.1) is satisfied.

$$\sigma^2_{B(j)} \leq \alpha\mu_T \tag{4.1}$$

where α is a pre-defined constant; μ_T is the average pixel value of the input image; and $\sigma^2_{B(j)}$ is computed using Equation (4.2), representing the weighted difference between the average pixel values of the two classes (T_0 and T_1) at the iteration j.

$$\sigma^2_{B(j)} = \omega_0\omega_1(\mu_1 - \mu_0)^2 \tag{4.2}$$

$$\text{where } \mu_0 = \frac{\mu_t}{\omega_0} \text{ and } \mu_1 = \frac{\mu_T - \mu_t}{\omega_1}$$

For each iteration, ω_0 represents the ratio of the pixels in class T_0 to the total number of pixels in the input image, and $\omega_1 = 1 - \omega_0$. The weighted means of the pixel values of the classes T_0 and T_1 are μ_0 and μ_1, respectively. They are computed from μ_t denoting the average pixel value of all the pixels in class T_0 in this iteration. Figure 4.3 (a) shows the original lumen image and Figure 4.3 (b) shows the initial lumen region obtained the above method.

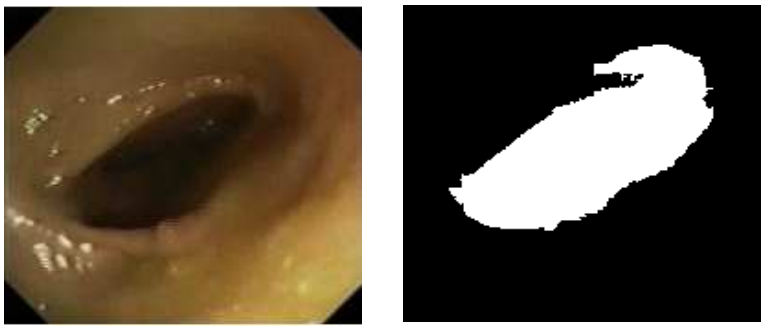

Figure 4.3. (a) Lumen Image, (b) Initial Lumen Region based on Otsu Thresholding.

4.2.2 Ellipse Fitting

Using the boundary of the initial lumen region, we generate the ellipses using the ellipse fitting method. An ellipse is be described by a second order polynomial [49] as follows:

$$F_{\mathbf{a}}(\mathbf{x}) = \mathbf{x} \cdot \mathbf{a} = ax^2 + bxy + cy^2 + dx + ey + f = 0 \tag{4.3}$$

with an equality ellipse-specific constraint

$$4ac - b^2 = 1 \tag{4.4}$$

where $\mathbf{a} = [a, b, c, d, e, f]^T$ is a set of coefficients in the ellipse, and $\mathbf{x} = [x^2, xy, y^2, x, y, 1]$ for (x, y) which is a set of coordinates of points lying on it. The polynomial $F_{\mathbf{a}}(\mathbf{x})$ is called the algebraic distance of the point (x, y) to the given conic. The fitting of an ellipse to a set of points (x_i, y_i), $i = 1, \cdots, N$ could be obtained by minimizing the sum of squared algebraic distances of the points considering Equation (4.4) as follows:

$$\min_a \Sigma_{i=1}^N F(x_i, y_i)^2 = \min_a \Sigma_{i=1}^N (\mathbf{x}_i \cdot \mathbf{a})^2 \tag{4.5}$$

We can obtain a set of coefficients ($\mathbf{a}$) by solving Equation (4.5) using the least square fitting method which was proposed in [49] as follows.

The ellipse-specific fitting problem (Equation (4.5)) can be formulated in the matrix form as

$$\min_a \parallel \mathbf{Da} \parallel^2 \text{ subject to } \mathbf{a}^T\mathbf{Ca} = 1 \tag{4.6}$$

where the matrix $\mathbf{D}$ is

$$\mathbf{D} = \begin{pmatrix} x_1^2 & x_1y_1 & y_1^2 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_i^2 & x_iy_i & y_i^2 & x_i & y_i & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N^2 & x_Ny_N & y_N^2 & x_N & y_N & 1 \end{pmatrix} \tag{4.7}$$

and the constraint matrix $\mathbf{C}$ expressing the constraint in Equation (4.4) is

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{4.8}$$

The minimization problem of Equation (4.6) can be solved by a quadratically constrained least square minimization proposed in [50]. By applying the Lagrange multipliers (λ), we get the following conditions for the optimal solution $\mathbf{a}$

$$\begin{aligned} \mathbf{Sa} &= \lambda \mathbf{Ca} \\ \mathbf{a}^T \mathbf{Ca} &= 1 \end{aligned} \tag{4.9}$$

where S is the scatter matrix of the size 6x6

$$\begin{aligned} \mathbf{S} &= \mathbf{D}^T \mathbf{D} \\ &= \begin{pmatrix} S_{x^4} & S_{x^3y} & S_{x^2y^2} & S_{x^3} & S_{x^2y} & S_{x^2} \\ S_{x^3y} & S_{x^2y^2} & S_{xy^3} & S_{x^2y} & S_{xy^2} & S_{xy} \\ S_{x^2y^2} & S_{xy^3} & S_{y^4} & S_{xy^2} & S_{y^3} & S_{y^2} \\ S_{x^3} & S_{x^2y} & S_{xy^2} & S_{x^2} & S_{xy} & S_x \\ S_{x^2y} & S_{xy^2} & S_{y^3} & S_{xy} & S_{y^2} & S_y \\ S_{x^2} & S_{xy} & S_{y^2} & S_x & S_y & S_1 \end{pmatrix} \end{aligned} \tag{4.10}$$

Due to the special structures of matrices $\mathbf{S}$ and $\mathbf{C}$, we can decomposed the matrices as follows. First, the matrix $\mathbf{D}$ can be decomposed into its quadratic and linear parts:

$$\mathbf{D} = (\mathbf{D}_1 \mathbf{D}_2) \tag{4.11}$$

where

$$\mathbf{D}_1 = \begin{pmatrix} x_1^2 & x_1y_1 & y_1^2 \\ \vdots & \vdots & \vdots \\ x_i^2 & x_iy_i & y_i^2 \\ \vdots & \vdots & \vdots \\ x_N^2 & x_Ny_N & y_N^2 \end{pmatrix}, \mathbf{D}_2 = \begin{pmatrix} x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots \\ x_i & y_i & 1 \\ \vdots & \vdots & \vdots \\ x_N & y_N & 1 \end{pmatrix} \tag{4.12}$$

The matrix $\mathbf{C}$ and matrix $\mathbf{S}$ can be split as follows:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \mathbf{S} = \begin{pmatrix} \mathbf{S}_1 & \mathbf{S}_2 \\ \mathbf{S}_2^T & \mathbf{S}_3 \end{pmatrix} \tag{4.13}$$

where

$$\mathbf{C}_1 = \begin{pmatrix} 0 & 0 & 2 \\ 0 & -1 & 0 \\ 2 & 0 & 0 \end{pmatrix} \text{ and } \begin{cases} \mathbf{S}_1 = \mathbf{D}_1^T\mathbf{D}_1 \\ \mathbf{S}_2 = \mathbf{D}_1^T\mathbf{D}_2 \\ \mathbf{S}_3 = \mathbf{D}_2^T\mathbf{D}_2 \end{cases}$$

Finally, we split the vector of coefficients $\mathbf{a}$ into

$$\mathbf{a} = \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} \text{ where } \mathbf{a}_1 = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \text{ and } \mathbf{a}_2 = \begin{pmatrix} d \\ e \\ f \end{pmatrix} \tag{4.14}$$

Based on these decompositions, Equation (4.9) can be rewritten as

$$\begin{pmatrix} \mathbf{S}_1 & \mathbf{S}_2 \\ \mathbf{S}_2^T & \mathbf{S}_3 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} = \lambda \cdot \begin{pmatrix} \mathbf{C}_1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} \tag{4.15}$$

which is equivalent to the following equations

$$\begin{cases} \mathbf{S}_1\mathbf{a}_1 + \mathbf{S}_2\mathbf{a}_2 = \lambda\mathbf{C}_1\mathbf{a}_1 \\ \mathbf{S}_2^T\mathbf{a}_1 + \mathbf{S}_3\mathbf{a}_2 = 0 \end{cases} \tag{4.16}$$

Regarding all the decomposition steps, Equation (4.9) can be reduced as follows.

$$\begin{aligned} \mathbf{M}\mathbf{a}_1 &= \lambda\mathbf{C}_1\mathbf{a}_1 \\ \mathbf{a}_1^T\mathbf{C}_1\mathbf{a}_1 &= 1 \end{aligned} \tag{4.17}$$

where M=$\mathbf{C}_1^{-1}(\mathbf{S}_1 - \mathbf{S}_2\mathbf{S}_3^{-1}\mathbf{S}_2^T)$ is the reduced scatter matrix of the size 3x3. Equation (4.17) is solved by using generalized eigenvectors such that

$$\| \mathbf{M}\mathbf{a}_1 \|^2 = \mathbf{a}_1^T\mathbf{M}^T\mathbf{M}\mathbf{a}_1 = \lambda \tag{4.18}$$

There are up to three real solutions $(\lambda^j, \mathbf{a}_1^j)$, but we can find one solution by looking for the eigenvector $\mathbf{a}_1^k$ corresponding to the minimal positive eigenvalue λ^k. Using the solution of $\mathbf{a}_1$, we can get $\mathbf{a}_2$ because $\mathbf{a}_2 = -\mathbf{S}_3^{-1}\mathbf{S}_2^T\mathbf{a}_1$. Finally, we can get all coefficients in $\mathbf{a}$ using Equation (4.14). Figure 4.4 (a) and (b) are from Figure 4.3 (a) and (b) respectively. Figure 4.4 (c) shows the boundary of Figure 4.4 (b) and Figure 4.4 (d) the shows the detected ellipse for lumen region.

Figure 4.4. (a) Lumen Image, (b) Initial Lumen Region, (c) Boundary of (b), and (d) Lumen Ellipse .

4.3 Experimental Results

In this section, we evaluate the effectiveness of the lumen identification algorithm. We test 3 colonoscopy videos and total number of frame is 3270 frames. The details are shown in (Table 4.1). The column "ID" represents the unique id number for each colonoscopy video, and the column "Total Frames" represents the total number of frames for each video. The column "Lumen view" represents the number of frames which have lumen regions and the column "Wall view" represents the number of frames which do not

have lumen regions for each video. This test set dose not contain any non-informative frames and one frame (instead of 30 frames/sec) is extracted per second for the evaluation.

Table 4.1. Test Set for Lumen Identification

ID	Total Frames	Lumen view	Wall view
009	1160	913	247
017	784	594	190
024	1573	1098	475

The parameters used in this evaluation includes the size threshold of 1500 pixels, the intensity threshold of 128, and the intensity difference threshold of 175. The experiment results are presented in Table 4.2. Column "S" represents the number of actual lumen frames we manually determined; column "T" represents the number of detected lumen frames; and column "C" represents the number of correctly detected lumen frames. Two performance metrics, recall and precision, are used. Table 4.2 shows high recall and precision for our proposed algorithm.

Table 4.2. Effectiveness of Lumen Identification

ID	S	T	C	Precision($\frac{C}{T}$)	Recall ($\frac{C}{S}$)
009	913	921	840	0.912	0.921
017	594	611	564	0.923	0.950
024	1098	1467	1297	0.884	0.908
Ave				0.907	0.926

CHAPTER 5

POLYP DETECTION

One of the most important tasks during the colonoscopy is to find polyps and early cancers. For computer-based detection of polyps of the stomach and colon, several techniques have been proposed using texture features [14, 15, 16, 51]. For instance, color wavelet covariance was used to generate a set of 72 texture features [14]. Using 180 training and 1200 testing images, the effectiveness of the color wavelet covariance was evaluated for the following questions: 1) which texture features among the 72 texture features are the most (or least) correlated with the presence of polyps, 2) what is the optimal color space, and 3) what is the effective window size? A polyp detection technique using Texture Spectrum and Neural Network classifier method has also been proposed [15]. However, a very limited number of images (54 abnormal images and 12 normal images) was tested in this study. More recently, the effectiveness of four different texture feature methods such as Texture Spectrum, Texture Spectrum with Color Histogram, Local Binary Pattern and Color Wavelet Covariance for detecting polyps were compared [16]. In this study, the Gaussian kernel Support Vector Machine (SVM) with 10-fold cross validation was used for the comparison of several texture features within 1000 selected images.

A major limitation of the above methods is that they depend on texture analysis with a fixed size window, and rely on their own training set of images for accuracy. Since even a single polyp can have different relative sizes and color features depending on the viewing position and distance of the camera from the polyp, it is not practical to use a fixed size window for texture analysis. For instance, Figure 5.1 shows an example of

a single polyp with apparent different sizes due to the different viewing position and distance between camera and polyp. For this reason, it is difficult to detect various sizes of polyps with one fixed window size.

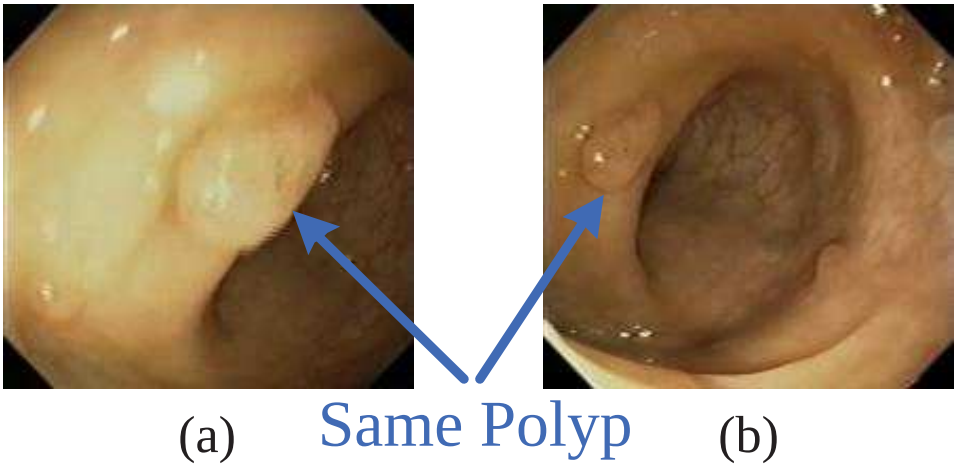

Figure 5.1. Same Polyp with Different Sizes: (a) ID-4920, and (b) ID-4995.

The other problem related to fixed-size window texture analysis is over-reflected areas in some polyps. An over-reflected area is seen when the light bundle hits wet mucosa surface perpendicular to the viewing direction, and is reflected straight back to the camera in the head of the endoscope. Such reflection significantly affects the texture analysis since an over-reflected area does not have the original surface information; some polyps display these areas, but others do not. An example of this can be seen in Figure 5.2. Figure 5.2 (a) shows an image of a polyp with an over-reflected area; Figure 5.2 (b) shows a window without and Figure 5.2 (c) shows a window with the over-reflected area in this polyp.

To overcome the above problems, we here propose a new technique that uses *shape* rather than texture. As seen in Figure 5.3, the shapes of polyps are 3D spherical or

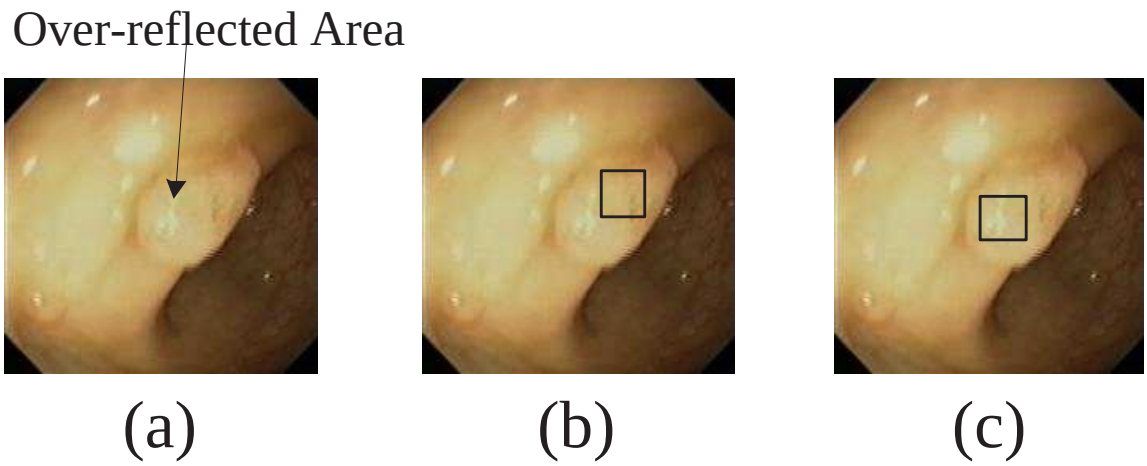

Figure 5.2. Over-reflected Area Problem: (a) Original Image, (b) A Window without Over-reflected Area in the Same Polyp, and (c) A Window with Over-reflected Area in the Same Polyp .

hemispherical forms in most cases so that polyps are represented as elliptical shapes in 2D images.

The method of detecting polyps using the elliptical shape has been proposed for CT colography [19, 20]. Figure 5.4 shows the general process of polyp detection methods using the elliptical shape. First, the edge lines are generated by thresholding the gradient magnitude of a CT image (Figure 5.4 (b)). Then, a line separating the maximal curvatures is detected (Figure 5.4 (c)). Using the detected curve line, the most similar ellipse is detected (Figure 5.4 (d)), and it is characterized by computing several curvatures. However, the polyp detection method for CT colography cannot be applied to colonoscopy images due to the following reasons.

- Unlike CT colography where X-ray density is the key discriminating factor, colonoscopy has to rely on color, relative difference of reflected light and shadowing as discriminating factors that result in detectable edges. And unlike CT colography where 3D reconstruction is readily achieved without loss of resolution, especially using the new 64-slice CT scanners, endoscopy at present has to rely on 2D information.

Figure 5.3. Examples of Polyp Shape .

- CT images are very different from colonoscopy images (Compare Figures 5.4 (a) and 5.5 (a)). In CT images, the colon wall can be easily separated from the lumen area (high contrast ratio) and a clear curve of the polyp protruding into the lumen typically exists (see Figure 5.4 (b)). However, as seen in Figure 5.5 (b), colonoscopy images contain more complicated edge lines so it is more difficult to group the edge lines that define a polyp.
- In CT images, general edge detection methods (Sobel edge detection, Canny edge detection, etc.) can be used because the boundary between colon wall and lumen is clear. For colonoscopy images we need a different method to obtain clear edges.
- As long as polyps are not flat, it is easy to segment polyp edges from the detected edge lines in CT images because 1) the edge line in a CT image is simple and 2) polyps in a CT image most often protrude from the colon wall into the lumen (Figure 5.4). The edge line in colonoscopy images is very complex and the assumption of the protuberance of polyp toward the lumen is not always true (flat lesions) or

appreciated at common angles of view (Figure 5.5). To segment polyp edges from detected edge lines, we need to use a different region segmentation method.

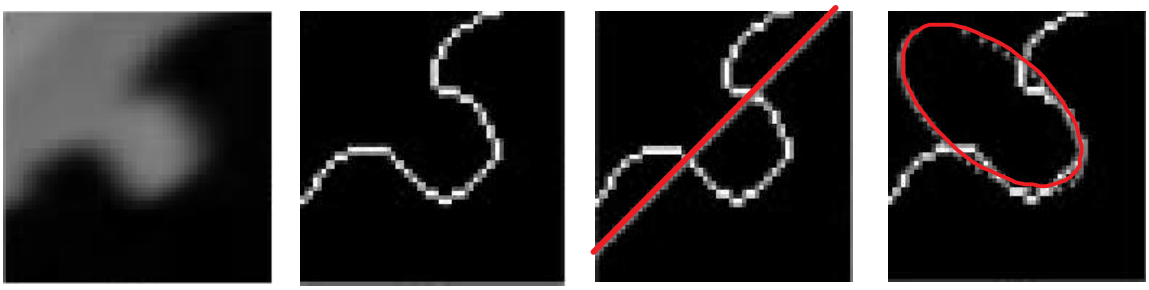

Figure 5.4. (a) Original CT Image, (b) Edge Line of (a), (c) Maximum Curve Separation from (b), and (d) Ellipse Detection.

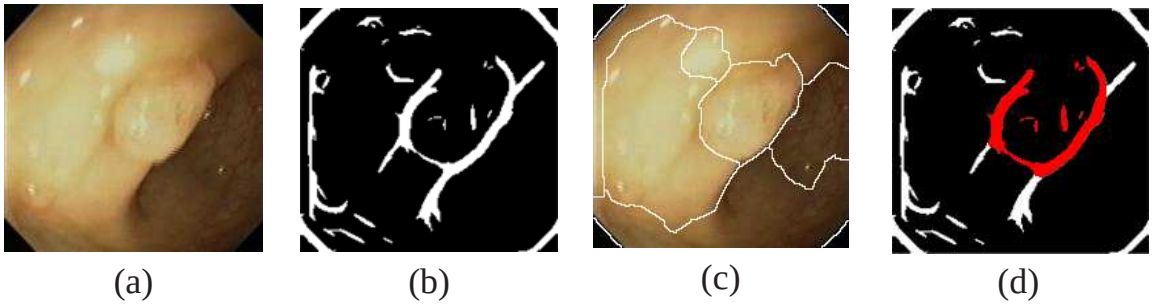

Figure 5.5. (a) Original Colonoscopy Image, (b) Edge Line of (a), (c) Segmented Regions of (a), and (d) Desirable Edges of Polyp.

In this paper, we propose a polyp region detection method based on the elliptical shape of polyps. Figure 5.6 shows the procedure of our proposed method.

First, a gradient magnitude is constructed using the matched filter method to get clear edge information. This gradient magnitude is used to generate the strong edge map based on the thresholding method and it is also used as the relief function of the watershed algorithm in the region segmentation step. In step 2, an image is

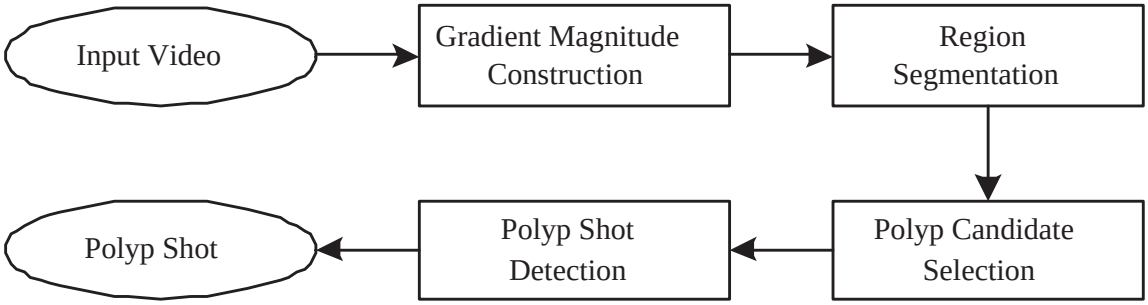

Figure 5.6. Polyp Detection Method.

segmented into several regions based on the marker-controlled watershed algorithm. In step 3, ellipses are detected using the ellipse fitting method using the information from the region segmentation and the strong edge. Among all detected ellipses, we select several ellipses which may represent polyps by the polyp candidate selection technique. Finally, the polyp shot detection step detects the missed polyp frames and determines the boundaries of polyp shots by utilizing an image registration technique. We note that the purpose of the 'polyp candidate' selection in Section 5.3 is to reduce the number of false polyp frames (i.e. the frames which are actually non-polyp frames but detected as polyp frames by the system) by finding the very obvious polyp as the polyp candidate. Base on the selected polyp candidate as seeds, the polyp shot detection technique discovers the missed polyp frames by comparing the polyp candidate frames with the remaining frames.

5.1 Gradient Magnitude Construction

To obtain clear edge information, a robust image gradient needs to be computed. Depending on the view point, many objects will have low contrast boundaries in colonoscopy video frames. To obtain robust gradient information, we use a matched filter method because this method can efficiently detect subtle as well as clear boundaries. The matched filter method has been used successfully to detect object boundaries within the retina

to define blood vessels [52, 53]. The matched filter method finds the maximum response among a family of filters. Figure 5.1 shows the basic idea of the matched filter method (i.e. a set of filter family is applied to a pixel (p) and find a filter with the maximum response value among a set of filters).

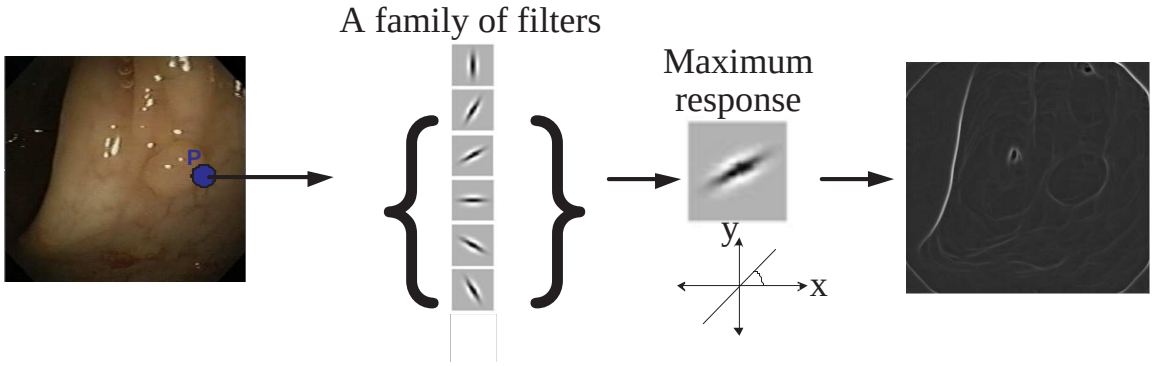

Figure 5.7. Matched Filtering.

Due to the characteristics of colonoscopy videos, there are many types of noise in the original frames that create edges. Examples of edge producing noises are light reflection and stool. Therefore it is necessary to reduce as much as possible the various noises in order to prevent false edge detection when the matched filter method is applied. To reduce noises while preserving the object boundaries, a median filter is applied because it performs well at removing noises and introduces very little blurring of edges [33]. To reduce the noises while preserving the object boundaries, a non-linear filter is applied because it performs well at removing noises from an image and causes blurring of edges very little [33]. We utilize a non-linear filter (i.e., median filter) for noise removal, and a noise reduced image ($\bar{I}$) will be computed as follows

$$\bar{I}(x, y) = I(x, y) \star MF \tag{5.1}$$

where I is the original image, MF is a median filter and $\star$ represents an ordering operation.

After reducing the noises, the matched filter method is applied to a noise-reduced image. The matched filter method needs to have a set of proper filters. We use the second derivative of two-dimensional Gaussian (GD^2) as a set of filters because GD^2 has two main advantages. First, even though it is known that GD^2 is sensitive to noises, we have found that GD^2 shows better performance than the first derivative of Gaussian (GD^1) in detecting edges in colonoscopy images. This is due to the fact that GD^2 can detect more detail edge information. The other advantage is that the line filter using GD^2 can take into account the direction of an edge so it works better when the boundaries are indistinct. A two-dimensional isotropic Gaussian function is defined as

$$G_\sigma(x,y) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp(-\frac{x^2+y^2}{2\sigma^2}) \tag{5.2}$$

The second derivative of two-dimensional Gaussian function along the x-axis is given by

$$G^2_\sigma(x,y) = \frac{\partial^2 G_\sigma(x,y)}{\partial x^2} = \frac{x^2-\sigma^2}{\sigma^4} G_\sigma(x,y) \tag{5.3}$$

By rotating the second derivative of Gaussian, a family of the second derivative of Gaussian along different orientations $\theta(0 \le \theta < \pi)$ can be generated as

$$\begin{aligned} G^2_{\sigma,\theta}(x,y) &= G^2_\sigma(x',y') \\ x' &= x\cos\theta + y\sin\theta \\ y' &= -x\sin\theta + y\cos\theta \end{aligned} \tag{5.4}$$

The gradient magnitude (GM) along the orientation θ, $GM(x,y,\theta)$, is defined as [33]:

$$GM(x,y,\theta) = \bar{I}(x,y) * G^2_{\sigma,\theta}(x,y) \tag{5.5}$$

where $\bar{I}$ is a noise-reduced image and * represents convolution. Equation (5.5) is defined for the gray level images but our colonoscopy images are in color, therefore, we extend it to color space. The gradient magnitude for a color image (GM_C) can be defined as:

$$GM_C = \max(GM_R, GM_G, GM_B) \tag{5.6}$$

where GM_R, GM_G, and GM_B are the gradient magnitudes for three color bands (R, G, B), respectively.

Using the gradient magnitudes, the strong edge information is obtained using the threshholidng method as follows. Let p be a pixel of the gradient magnitude (GM_c). Then the strong edge map (B) can be obtaining by assigning '1' if p is larger than a certain threshold (TH_{eg}), otherwise '0' is assigned. Figure 5.8 (a) shows the noise reduced image. Figure 5.8 (a) shows the color gradient map of (a) and Figure 5.8 (c) shows the strong edge obtained from (b).

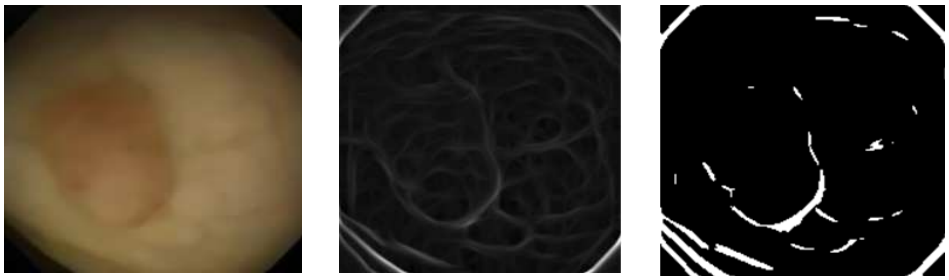

Figure 5.8. (a) Noise Reduced Image ($\bar{I}$), (b) Gradient Magnitude of Color Image (GM_C), and (c) Strong Edge (B) .

5.2 Region Segmentation

In this section, we propose a region segmentation method based on the marker-controlled watershed algorithm. Despite the popular usage of the watershed algorithm

in the literature of the image segmentation field, the selection of the initial marks is still open problems. In this section, we propose a novel mark selection method using the strong edge information.

5.2.1 Watershed Algorithm

Region segmentation is a fundamental step in image analysis. It should yield a partitioning of an image into disjoint regions. The segmentation process can rely on the uniformity of the feature within the regions or on the edge evidence. As seen in Figure 5.9, a typical polyp consists of different color values (or intensity values) that in part depend on the relative distance between the polyp surface and the light source. This distance varies because the shape of a polyp is 3D sphericity. Thus, a polyp could be recognized by its edge evidence not by the region uniformity. However, depending on the light condition and viewing position, only some parts of a polyp boundary have strong edge information and others have weak edge information.

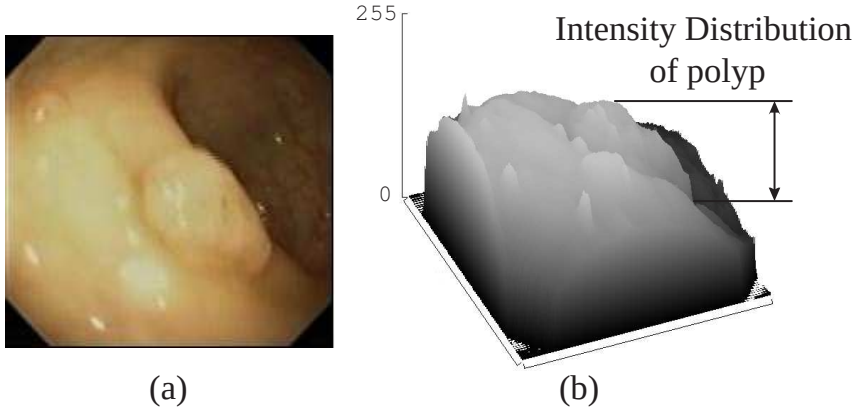

Figure 5.9. (a) Polyp Image, and 3D Intensity Value Plot of (a).

Based on these observations, we use the marker-controlled watershed algorithm for polyp segmentation because it has some advantages [54, 55]: the watershed algorithm

has been proven powerful for contour detection as well as region segmentation because the watershed lines always correspond to the most significant edges between markers. Thus, the watershed algorithm is not affected by lower-contrast edges due to noises. In addition, even if there are no strong edges between the markers, the watershed algorithm always detects a contour in an area. Of importance, it can handle the gap between broken edges properly, and place the boundaries at the most significant edges.

The watershed algorithm uses a topological relief function representing edge evidence as input. For the relief function, we use the gradient magnitude based on the matched filter method obtained in Section 5.1. By viewing this function as a mountain landscape, object boundaries are determined as watershed lines. The watershed algorithm selects small number of markers as the initial seeds. The segmentation result is directly related with the number and the position of the markers. Most of segmentation algorithms based on watershed transform [56, 57, 58] select the markers by computing the significant local minima or local maxima of the gradient magnitude. Using the selected markers, the watershed transform is performed. As the relief goes deeper into the water, the regions surrounding the seeds become flooded, that is, the catchment basins are constructed. When two or more regions merged to a point, a dam (i.e. watershed line) is raised. Eventually, the whole image will be partitioned into catchment basins which are bordered by the watershed lines. The number of catchment basins cannot be different from the number of seeds.

5.2.1.1 Marker Selection

The quality of region segmentation results is determined by the proper number of marks (i.e. one mark for one polyp) because the number of segmented regions should be the same as the number of seeds.

The shapes of polyps are 3D spherical or hemispherical forms so the geometric elevation level (i.e. light intensity) of a polyp is higher than the surrounding regions in colonoscopy images. Using this characteristic, we select regions which have higher intensity values than surrounding regions as initial markers by using the morphological operation called *regional maxima* [59]. Before applying the regional maxima to the original image, we need to reduce the effect from unreliable high intensity pixels. The effect from unreliable high intensity pixels can be reduced by utilizing the morphological reconstruction operation [59]. The morphological reconstruction, $\rho_I(J)$, is defined as follows:

$$\rho_I(J) = \bigvee_{n \geq 1} \delta_I^{(n)}(J) \tag{5.7}$$

where $\delta_I^{(n)}$ is the grayscale geodesic dilation of size n, J is the marker image and I is the mask image. Both J and I are identical in size and $J \leq I$ (i.e. a pixel value in J is less than or equal to the corresponding pixel value in I). The geodesic dilation of size n ($\delta_I^{(n)}$) is defined as follows:

$$\delta_I^{(n)} = \underbrace{\delta_I^{(1)} \circ \delta_I^{(1)} \circ \cdots \circ \delta_I^{(1)}}_{n\,times}(J) \tag{5.8}$$

And, the geodesic dilation of size 1 ($\delta_I^{(1)}$) is defined as follows:

$$\delta_I^{(1)} = (J \oplus B) \wedge I \tag{5.9}$$

where $\wedge$ stands for the pointwise minimum and $J \oplus B$ is the dilation of J by the 3×3 structuring element B.

Based on the definition of the reconstruction operation, we propose two sequential operations ('opening-by-reconstruction' followed by 'closing-by-reconstruction') to remove unreliable pixels as follows:

$$\begin{aligned} \text{opening-by-reconstruction: } I_{or} &= \rho_I(I \ominus S) \\ \text{closing-by-reconstruction: } I_{orcr} &= \rho_{I'_{or}}(I'_{or} \oplus S) \end{aligned} \tag{5.10}$$

where $\ominus S$ is the erosion by the element S and $\oplus S$ is the dilation by the element S. We use the 5×5 disk element for S. I'_{or} stands for the complement image of I_{or}. The complement image (I'_{or}) can be calculated by subtracting each pixel value of I_{or} from the maximum pixel value of I_{or}. Then, the markers are selected by finding regional maxima of I_{orcr}. A regional maximum is a connected component set of pixels with the value t whose external boundary pixels are strictly smaller than t. Formally, the regional maximum (M) of I_{orcr} can be defined as [59]:

$$\text{Regional Maximum } (M) \text{ at level } t \iff M \text{ connected and} \begin{cases} \forall p \in M & I_{orcr}(p) = t \\ \delta_M^{(1)} \backslash M & I_{orcr}(p) < t \end{cases} \tag{5.11}$$

where $\backslash$ is the set difference operation.

However, this technique generates too many markers causing an over-segmentation for colonoscopy image. As seen in Figure 5.10, there are two marks (M_6 and M_{15}) on a polyp region, which results in wrong segmentation dividing a polyp into two regions.

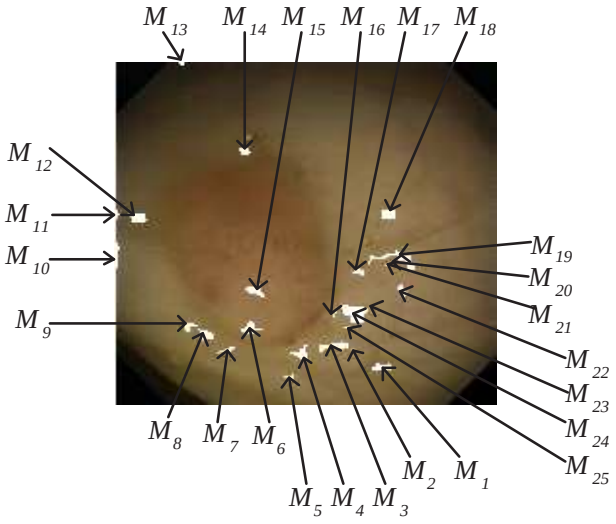

Figure 5.10. Initial Marks .

To prevent this over-segmentation problem, the marks M_6 and M_{15} need to be merged as one mark. The mark merging method are proposed based on the following two observations: (1) the marks on the same polyp are not far from one anther and (2) there is no strong edge information between the marks if the marks are obtained at the same polyp. For instance, Figure 5.11 (a) shows that M_7, M_8 and M_{15} are located within a certain searching distance (TH_r) from M_6, so they could be merged as one mark. However as seen in Figure 5.11 (b), there are strong edge information between M_6 and M_7, and between M_6 and M_8. In contrast, there is no strong edge information between M_6 and M_{15}. Thus, these two marks are connected as one mark by drawing a line between M_6 and M_{15}.

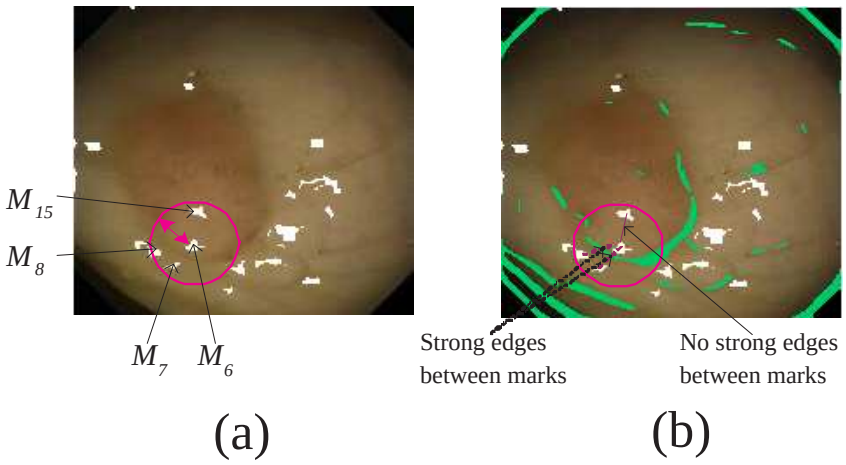

Figure 5.11. (a) Searching Area, (b)Marker Merging .

Using this method, the initial 25 marks are reduced in 9 marks as seen in Figure 5.12.

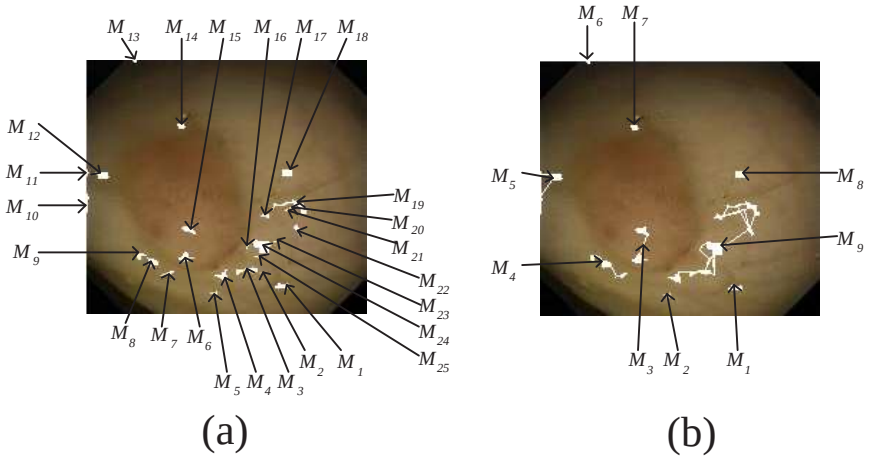

Figure 5.12. (a) Initial Markers, (b) Merged Markers .

5.3 Polyp Candidate Selection

In this section, we propose the polyp candidate selection method based on the elliptical shape of polyps.

5.3.1 Ellipse Fitting

Using the edges in each segmented region, we generate the ellipses using the ellipse fitting method. First, the binary edge map is constructed for each segmented region by defining the binary edge map (B_i) of the region i is the strong edge (B) in the region i. For instance, Figure 5.13 (b) is the strong edge information (B) presented in Section 5.1 , and Figure 5.13 (c) is the binary edge map (B_3) for the region 3.

Using the binary edge map of each region, we will find the best ellipse using the least square fitting method, which is explained in Section 4.2.2. Figure 5.14 (a) and (b) are from Figure 5.13 (a) and (b) respectively. Figure 5.14 (c) shows the detected ellipse for each region using the above algorithm. We note that ellipses are not detected for the regions 1 (R_1), 2 (R_2), 5 (R_5), 6 (R_6), 7 (R_7), and 8 (R_8) because there is no edge information in the binary edge maps corresponding to these regions.

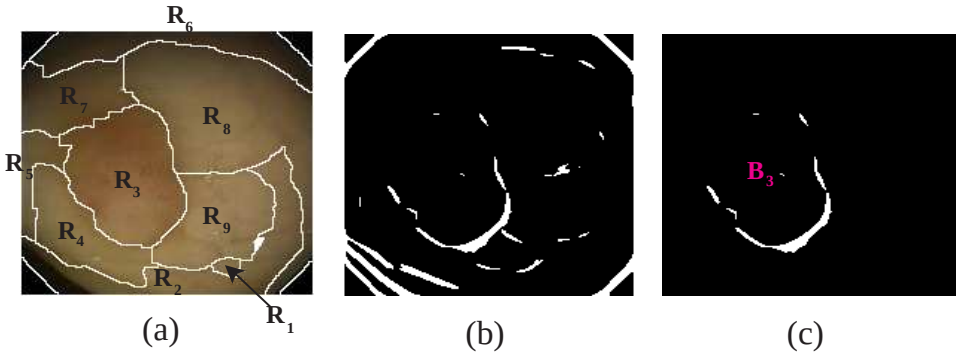

Figure 5.13. (a) Segmented Regions, (b) Strong Edge, and (c) Binary Edge Map of Region 3 .

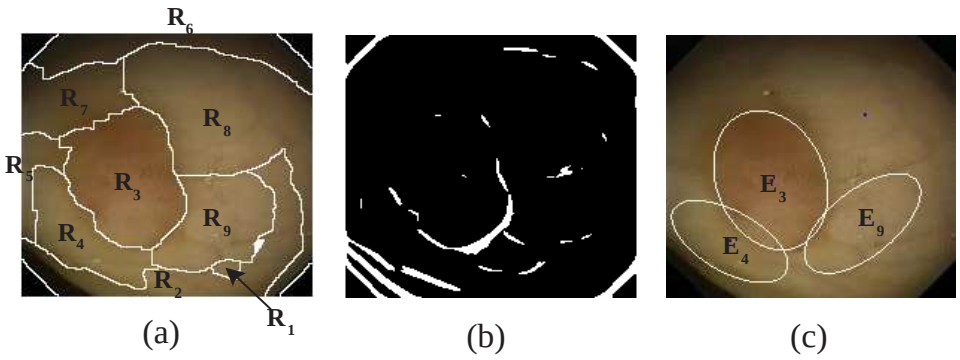

Figure 5.14. (a) Segmented Regions, (b) Binary Edge Map, and (c) Detected Ellipses .

Among the detected ellipses, we need to select the ellipses which have a possibility to represent polyps, and remove the ellipses which are not polyps. In this section, we propose three filtering methods: filtering by curve direction and semimajor-semiminor ratio, filtering by edge distance, and filtering by intensity value. After filtering out the irrelevant ellipses by the proposed method, we declare a frame as a polyp candidate frame if it has any remaining ellipses.

5.3.2 Filtering by Curve Direction and Semimajor-Semiminor Ratio

Different edge shapes can generate similar ellipse. For instance, Figure 5.15 (a) shows two different edge maps in which the upper edge map is obtained from a polyp

frame and the lower edge map is obtained from a non-polyp frame. As seen in Figure 5.15 (b), the same ellipses can be generated from two different frames. Thus, we need to compare edge shape in edge map with that in ellipse to distinguish the polyp ellipses from the non-polyp ellipses. Figure 5.15 (c) shows the best fitting parabolas obtained using the parts A and C of Figure 5.15 (a), which are indicated with red rectangles. Figure 5.15 (d) shows the best fitting parabolas obtained using the parts $\overline{A}$ and $\overline{C}$ of Figure 5.15 (b), which are indicated with blue rectangles. The arrows in Figure 5.15 (c) and (d) represent the direction of parabolas. If an ellipse is generated from a polyp, the direction of the parabola from any part of ellipse and the direction of the parabola from the corresponding part of edges are same (the top images of Figure 5.15 (c) and (d)). In contrast, if there is a ceratin part in which the direction of the parabola from an ellipse and the direction of the parabola from the corresponding edges are different (the bottom image of Figure 5.15 (c) and (d)), then the ellipse is not generated from a polyp.

Based on this observation, we propose to divide edges into four parts and calculate the curve information for each part as follows. The ellipse represented in Equation (4.3) can be expressed as follows:

$$\frac{(x'-c_x)^2}{a^2}+\frac{(y'-c_y)^2}{b^2}=1 \tag{5.12}$$
$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

where (x, y) are coordinates of points, a is the length of the semimajor axis, b is the length of the semiminor axis, and (c_x, c_y) is the center point of the ellipse. As seen in Figure 5.16 (a), two foci (F_1 and F_2) can be easily calculated because $a^2 = b^2 + c^2$. Illustrated in Figure 5.16 (b), we divide an ellipse into four parts based on the two foci points (F_1 and F_2): (1)-upper side of the line between F_1 and F_2, (2)-right side of F_2, (3)-lower side of the line between F_1 and F_2, and (4)-left side of F_1. By selecting the

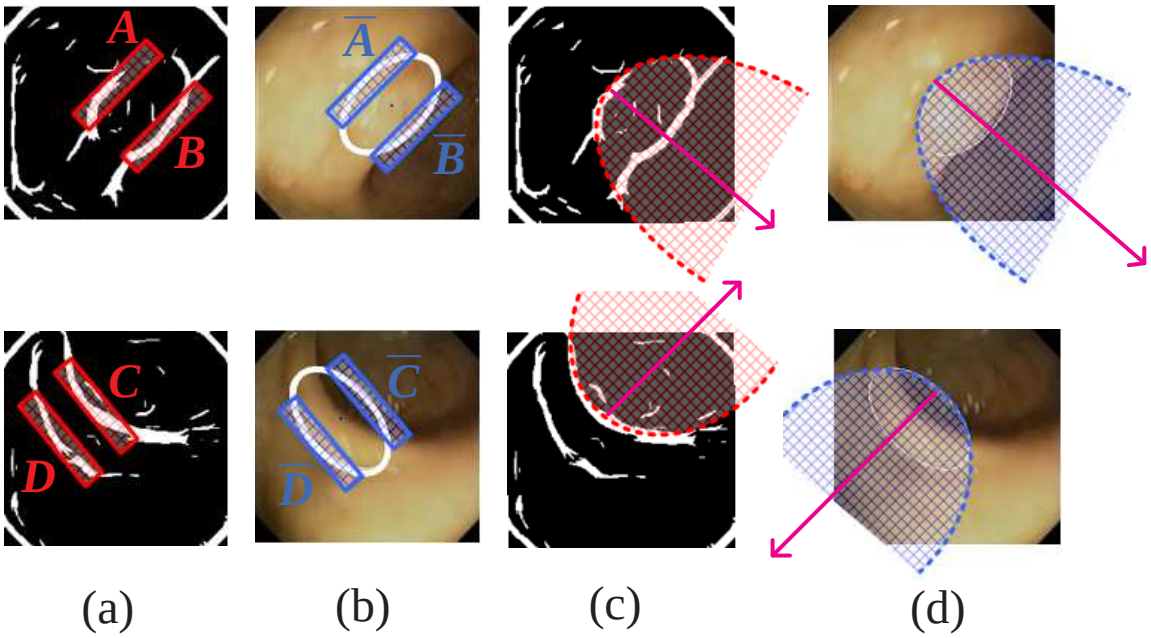

Figure 5.15. (a) Binary Edge Maps, (b) Detected Ellipses from (a), (c) Parabolas generated from Parts A and C in (a), and (d) Parabolas generated from $\overline{A}$ and $\overline{C}$ in (b).

edges in the corresponding parts, we can divide the edges into four parts. We call a part i of an ellipse as a *dismembered-ellipse* (E^i) shown in Figure 5.16 (c), and a part i of edges as a *dismembered-edge set* (B^i) shown in Figure 5.16 (d).

For each dismembered-edge set (B^i, $i = 1, \cdots, 4$) , we compute the curve direction and the max curvature by detecting a parabola using the polynomial curve fitting method. The second order polynomial of a parabola is same as the second order polynomial of an ellipse (Equation (4.3)) but it has a different constraint ($b^2 - 4ac = 0$). It is known that the second order polynomial of a parabola cannot be solved using the least square fitting because the constraint of a parabola is $b^2 - 4ac = 0$ [60]. Therefore, we use another curve model for a parabola as follows [60]:

$$f(x) = \alpha + \beta x + \gamma x^2 \tag{5.13}$$

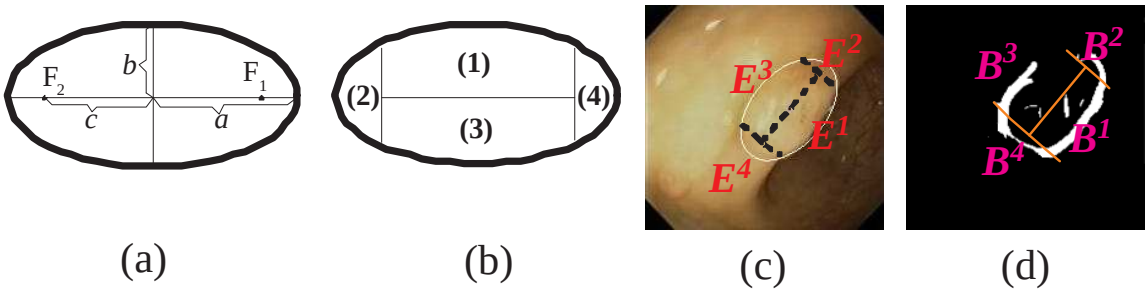

Figure 5.16. (a) Ellipse, (b) Ellipse with Four Part, (c) Dismembered-Ellipse, and (d) Dismembered-Edge Set .

Equation (5.13) can model a parabola only if the directrix of a parabola is parallel to the x-axis described in Figure 5.17 (a). We have to rotate a dismembered-edge set (B^i) if B^i is not in the proper position. We define θ ($0 < \theta \leq \pi$) as the counterclockwise angle from the x-axis to the major axis of an ellipse described in Figure 5.17 (b).

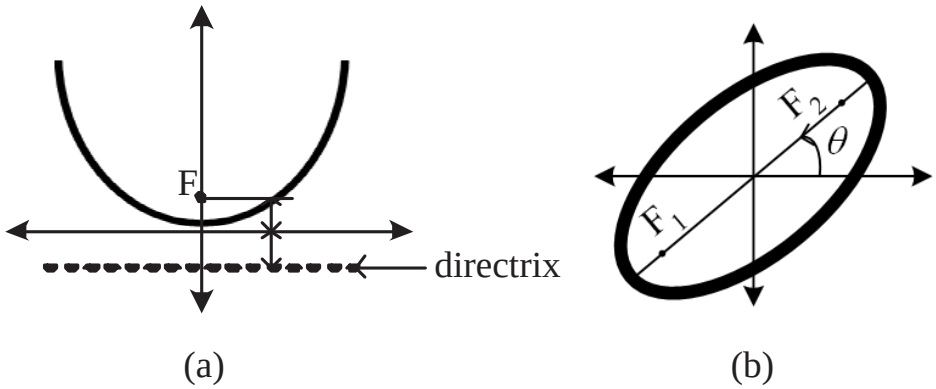

Figure 5.17. (a) Parabola parallel to x-axis, and (b) Counterclockwise Angle .

Based on θ, we can place each dismembered-edge set (B^i) to fit Equation (5.13) by rotating each dismembered-edge set (B^i, $i = 1, \cdots, 4$) by $\theta + \frac{\pi(i-1)}{2}$. Figure 5.18 (b), (c), (d) and (e) show the rotated B^i by θ, $\theta + \frac{\pi}{2}$, $\theta + \frac{2\pi}{2}$ and $\theta + \frac{3\pi}{2}$, and the fitted parabola f^i for the corresponding B^i, respectively.

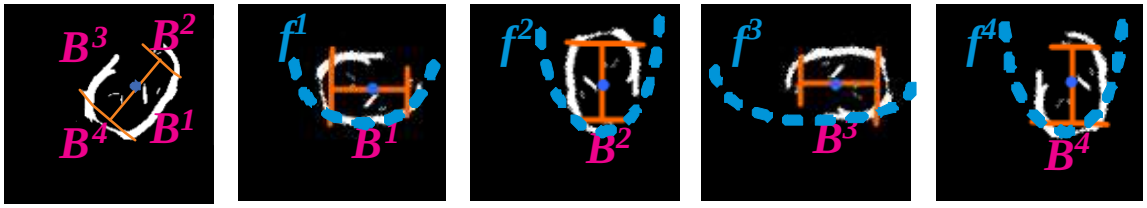

Figure 5.18. (a) Original Edge, (b) Rotated by θ, (c) Rotated by $\theta + \frac{\pi}{2}$, (d) Rotated by $\theta + \frac{2\pi}{2}$, and (e) Rotated by $\theta + \frac{3\pi}{2}$.

After rotating a dismembered-edge set (B^i), we can calculate the coefficients (α, β and γ) of Equation (5.13) as follows. Given a set of pixels ($p(x_i, y_i), i = 1, \cdots, n$) belonging to a dismembered-edge set (B^i), the coefficients of the second degree parabola can be obtained when the least square error (LSE) is minimized.

$$LSE = \sum_{i=1}^{n} [y_i - f(x_i)]^2 = \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i + \gamma x_i^2)]^2 \tag{5.14}$$

The condition for LSE to be minimized is the first derivatives of LSE to zero such as:

$$\begin{aligned} \frac{\partial LSE}{\partial \alpha} &= 2 \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i + \gamma x_i^2)] = 0 \\ \frac{\partial LSE}{\partial \beta} &= 2 \sum_{i=1}^{n} x_i [y_i - (\alpha + \beta x_i + \gamma x_i^2)] = 0 \\ \frac{\partial LSE}{\partial \gamma} &= 2 \sum_{i=1}^{n} x_i^2 [y_i - (\alpha + \beta x_i + \gamma x_i^2)] = 0 \end{aligned} \tag{5.15}$$

$$\frac{\partial LSE}{\partial \alpha} = 0, \ \frac{\partial LSE}{\partial \beta} = 0, \ \frac{\partial LSE}{\partial \gamma} = 0 \tag{5.16}$$

By minimizing the above equations, we can get the coefficients as follows:

$$\begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \end{pmatrix} \tag{5.17}$$

By applying the above polynomial curve fitting method to four dismembered-edge sets (B^i, $i = 1, \cdots, 4$) , we can obtain up to four curves: f^1, f^2, f^3 and f^4.

In addition, typical polyps have a certain ranges of semimajor-semiminor ratio ($\frac{a}{b}$). Based on the coefficients of parabolas and the semimajor-semiminor ratio, an ellipse (E) is declared as a polyp candidate if an ellipse satisfy both of the following two conditions. Otherwise, it is not a polyp and filtered out.

- Condition 1: If each dismembered-edge set is a part of a polyp, the coefficient γ^i of a parabola f^i should be larger than zero because the direction of f^i is turned up (see Figure 5.18). So, if there is a γ^i ($i = 1, \cdots, 4$) which is less than or equal to zero, the ellipse is not a polyp candidate.
- Condition 2: If semimajor-semiminor ratio ($\frac{a}{b}$) be a certain range ($1 \leq \frac{a}{b} \leq TH_{k1}$), then the ellipse is a polyp candidate. Otherwise, it is filtered out. We use 3 for TH_{k1}.

5.3.3 Filtering by Lumen

After filtering out the ellipses by the curve direction and semimajor-semiminor ratio, we examine the remaining ellipses by their color. As seen in Figure 5.19, a shape of lumen is an elliptical shape and it is detected along with strong edges.

Figure 5.19. (a) Original Edge, (b) Binary Edge Map, and (c) Detected Ellipse .

Lumen areas are easily misclassified as polyps. Even though a lumen is similar to a polyp in shape, it is different from in color (intensity) because a lumen is relatively darker, and a polyp is the relatively brighter. Using the lumen identification technique presented in Chapter 4, we filter out an ellipse if it is in the lumen region.

5.3.4 Filtering by Edge Distance

After filtering out the ellipses by two filtering methods, we evaluate the remaining ellipses with another polyp characteristic. Even though the entire boundary of polyp has not strong edge information, some parts of polyp boundary must have strong edge information along the detected ellipse. Figure 5.20 (a) shows the typical patterns of strong edges of polyps in the colonoscopy image.

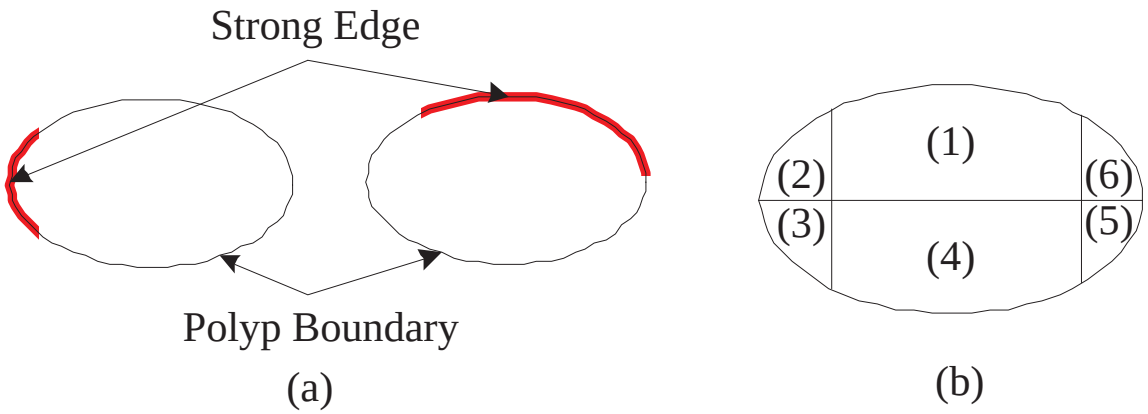

Figure 5.20. (a) Strong Edge Pattern of Polyp, and (b) Ellipse with Six Parts .

To characterize the above polyp edge patterns, we divide an ellipse into six parts as seen in Figure 5.20 (b) so we have six dismembered-ellipses (E^i, $i = 1, \cdot, 6$) and dismembered-edge sets (B^i, $i = 1, \cdot, 6$). By modifying the hausdorff distance [61], we

define the edge distance (ED) as the sum of the distance between a dismembered ellipse (E^i) and the corresponding dismembered edge set (B^i) as follows:

$$ED^i = ED(E^i, B^i) = \sum_{a \in E^i} \min_{b \in B^i} d(a, b) \tag{5.18}$$

where a and b are points of E^i and B^i respectively, and $d(a, b)$ is the Euclidian distance between a and b. ED measures how much a dismembered edge set is (dis)similar to the corresponding dismembered ellipse. The edge distance ED is asymmetric such as $ED(E^i, B^i) \neq ED(B^i, E^i)$, therefore, it can find if there are strong edges along the detected ellipse. Based on the edge distance (ED), an ellipse (E) is declared as a polyp candidate if either of the following conditions is satisfied. Otherwise, it is not a polyp.

- Condition 3: If there are strong edges close to an ellipse boundary in parts (2) and (3), or in parts (5) and (6), then the ellipse is a polyp candidate. It can be formulated as follows:

  $$\min(ED^{(2,3)}, ED^{(5,6)}) \leq TH_\zeta$$

- Condition 4: If there are strong edges close to an ellipse boundary in parts (1) and (2), or in parts (1) and (6), or in parts (3) and (4), or in parts (4) and (5), then the ellipse is a polyp candidate. It can be formulated as follows:

  $$\min(ED^{(1,2)}, ED^{(1,6)}, ED^{(3,4)}, ED^{(4,5)}) \leq TH_\xi$$

where $ED^{(u,v)} = ED^{(u)} + ED^{(v)}$.

5.4 Polyp Shot Detection

In this section, we propose the polyp shot detection method based on image registration to detect the missed polyp frames and determine the boundaries of polyp shots by comparing the polyp candidate frames with their adjacent frames.

For a polyp candidate frame, its adjacent frame is registered based on the mutual information method [62]. Given two frames A (a polyp candidate frame) and B (its adjacent frame), the definition of the Mutual Information $MI(A, B)$ of these frames is:

$$MI(A, B) = H(A) + H(B) - H(A, B) \tag{5.19}$$

where $H(A)$ and $H(B)$ are the entropies of the frames A and B, and $H(A, B)$ is their joint entropy. The definitions of these entropies are:

$$H(A) = -\sum_{a} P_A(a) \cdot \log P_A(a) \tag{5.20}$$

$$H(B) = -\sum_{b} P_B(b) \cdot \log P_B(b) \tag{5.21}$$

$$H(A, B) = -\sum_{a,b} P_{A,B}(a, b) \cdot \log P_{A,B}(a, b) \tag{5.22}$$

where $P_A(a) = \sum_{b} P_{A,B}(a, b)$ and $P_B(b) = \sum_{b} P_{A,B}(a, b)$ are the probabilities of histograms. $P_{A,B}(a, b) = \frac{h(a,b)}{\sum_{a,b} h(a, b)}$ is the joint probability and h is a joint histogram for the frame pair.

The MI registration criterion states that the highest value of the MI can be obtained when the frame pair is geometrically aligned through a geometric transformation (T). We use the rigid body transformation as our geometric transformation (T) and use the simplex method [62] to maximize the mutual information measure under the rigid body transformation. We note that we convert the color images into the gray-level images before the image registration is performed. Figure 5.21 (a) is a polyp candidate frame (A), and Figure 5.21 (b) is an adjacent frame (B). Figure 5.21 (c) is obtained by registering the adjacent frame (B) into the polyp candidate frame (A). Figure 5.21 (d) and (e) are the corresponding binary edge map of Figure 5.21 (a) and (b), respectively. Figure 5.21 (f) is obtained by transforming Figure 5.21 (e) using the same parameters of Figure 5.21 (c)

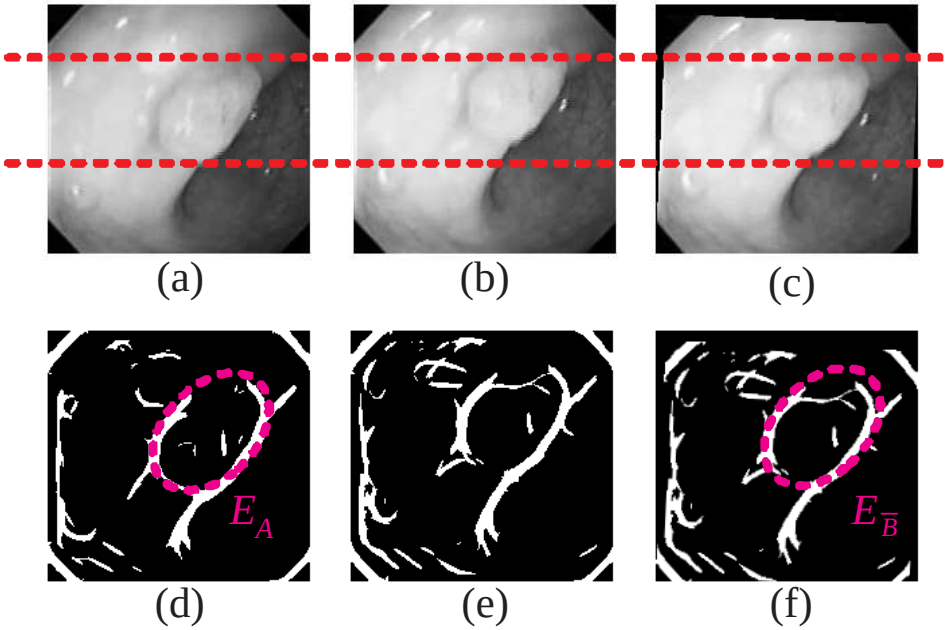

Figure 5.21. (a) Polyp Candidate Frame (A), (b) Adjacent Frame (B), (c) Registered Adjacent Frame ($\overline{B}$), (d) Binary Edge Map of (a), (e) Binary Edge Map of (b), and (f) Registered Adjacent Binary Edge Map.

After two frames (A and B) are registered, two edge sets (E_A and $E_{\overline{B}}$) are generated by selecting the edge pixels within the ellipse of polyp candidate frame (A) and the registered adjacent frame ($\overline{B}$). To examine if E_A and $E_{\overline{B}}$ have the similar edge pattern, we measure the distance ($Dist$) between E_A and $E_{\overline{B}}$ as follows:

$$Dist(E_A, E_{\overline{B}}) = \max(ED(E_A, E_{\overline{B}}), ED(E_{\overline{B}}, E_A)) \tag{5.23}$$

where ED is the edge distance which is defined in Section 5.3.4. If the $Dist(E_A, E_{\overline{B}})$ is less than a certain threshold TH_η, a polyp candidate frame(A) and its adjacent frame (B) have a same polyp. Otherwise, there is no polyp in the adjacent frame (B). As seen in Figure 5.22, a polyp shot is detected with four steps as follows:

- Step 1: Let A_i be a polyp candidate frame i and A_j be its left adjacent frame ($j = i - 1$). The registered adjacent frame $\overline{A_j}$ is generated using the mutual

information based image registration, and make two edge sets (E_{A_i} and $E_{\overline{A_j}}$) within an ellipse.

- Step 2: Measure a distance ($Dist$) between E_{A_i} and $E_{\overline{A_j}}$. If $Dist(E_{A_i}, E_{\overline{A_j}}) < TH_\eta$, set $i = i - 1$ to replace A_i with A_j, and set $j = j - 1$ to select the left adjacent frame of A_j for new A_j. Using the new assigned A_i and A_j, repeat Step 1. If $Dist(E_{A_i}, E_{\overline{A_j}}) \geq TH_\eta$, the left-side boundary of a polyp shot is declared and move to Step 3.
- Step 3: Repeat the same procedure in Step 1 and Step 2 to detect the right-side boundary of a polyp shot with the different adjacent frame (i.e. A_j, $(j = i + 1)$).
- Step 4: Count the number of polyp candidate frames in a shot. If the number of the polyp candidate frames is larger than a certain threshold (TH_τ), the shot is declared as a polyp shot.

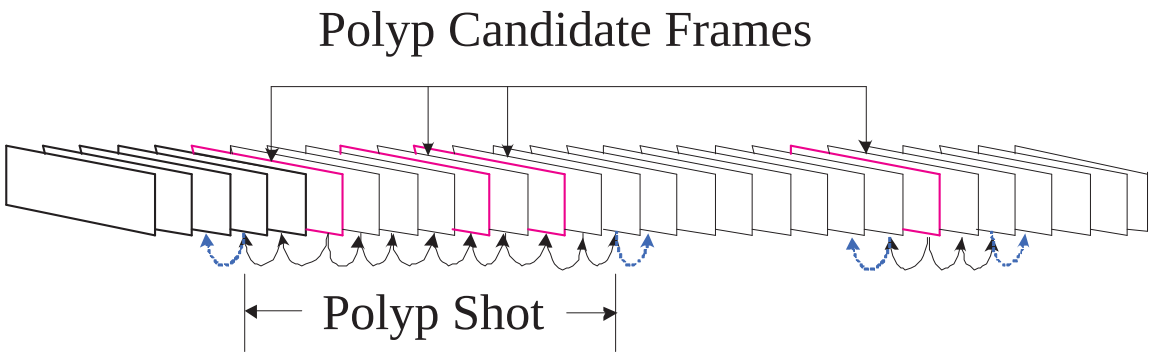

Figure 5.22. Process of Polyp Shot Detection.

5.5 Experimental Result

In this section, we assess the effectiveness of the proposed polyp detection technique. There are several threshold values in the marker selection, filtering by edge distance and

polyp shot detection. First, we show how to select the proper threshold values. After than, we evaluate the polyp detection algorithm with three colonoscopy videos.

5.5.1 Evaluation of Marker Selection

Our polyp detection algorithm uses the results of the region segmentation based on the watershed algorithm. The quality of the watershed algorithm is directly related with the number of markers. Only one marker should be selected a polyp. We propose the marker selection technique in Section 5.2.1.1 and there are two thresholds: threshold for the strong edge construction (TH_{eg}) and threshold for the searching distance (TH_r). We tested the marker selection technique using 400 polyps which is randomly selected. Table 5.1 shows the experimental results of our marker selection technique with different (TH_{eg}) values. The first column (TH_{eg}) represents the threshold to construct the strong edge map (B), and the second column shows the number of correctly selected polyp markers (i.e. a polyp has only one marker). The correct ratio is obtained by the number of correctly selected polyp markers divided by the total number of polyps (i.e. 400 polyps). This experiment is performed when the searching distance is 15 ($TH_r = 15$). Table 5.1 shows that we can select polyp markers correctly with very high accuracy (over 95%) when we use $TH_{eg} = 1.1$.

Table 5.2 shows the experimental results of our marker selection technique with different searching distance (TH_r) values. The first column (TH_r) represents the threshold of the searching distance, and the second column shows the number of correctly selected polyp markers (i.e. a polyp has only one marker). The correct ratio is obtained by the number of correctly selected polyp markers divided by the total number of polyps (i.e. 400 polyps). This experiment is performed when $TH_{eg} = 1.1$, and Table 5.1 shows that most of polyp markers can be correctly selected with $TH_r = 15$.

Table 5.1. Result of Polyp Maker Selection with Different TH_{eg}

TH_{eg}	# of Correct Polyp Marker	Correct Ratio
0.1	97	0.242
0.2	132	0.330
0.3	155	0.388
0.4	192	0.482
0.5	225	0.563
0.6	292	0.731
0.7	310	0.776
0.8	347	0.868
0.9	382	0.955
1.0	382	0.957
1.1	383	0.958
1.2	354	0.887
1.3	322	0.807
1.4	314	0.786
1.5	299	0.749

5.5.2 Evaluation of Filtering by Edge Distance

There are two thresholds related with the edge distance: TH_{ζ} of Condition 3 and TH_{ξ} of Condition 4 in Section 5.3.4. In this section, we shows how to select these two thresholds. For this experiment, we extract the entire frames of a colonoscopy video with 15 frames-per-second rate. The duration of the colonoscopy video is 5 minutes so we have 8,621 frames which consists of 815 polyp frames and 7,806 normal frames. The polyp

Table 5.2. Result of Polyp Maker Selection with Different TH_r

TH_r	# of Correct Marker Frames	Correct Ratio
5	279	0.698
10	333	0.833
15	383	0.958
20	379	0.949
25	340	0.850
30	323	0.808

frame represents a frame containing polyps, and the normal frame represents a frame in which no polyp is included. Table 5.3 shows the detail of our data set.

Table 5.3. Test Set

Class of Frame	# of Frame	Frame Size (Pixel)
Polyp Frames	815	195x187
Normal Frames	7806	195x187

Figure 5.23 shows that the four performance metrics of the polyp detection based on TH_ζ and TH_ξ, in which the x-axis represents the threshold values and the y-axis represents the values from 0 to 1. Figure 5.23 (a) shows the performance metrics of the polyp detection when only TH_ζ is used regarding different threshold values, and it shows the gradual increase in the sensitivity and the steep decrease in three other metrics after $TH_\zeta = 7$. Figure 5.23 (b) shows the performance metrics of the polyp detection when only TH_ξ is used regarding different threshold values, and it shows the gradual increase in the sensitivity and the steep decrease in three other metrics after $TH_\xi = 4$.

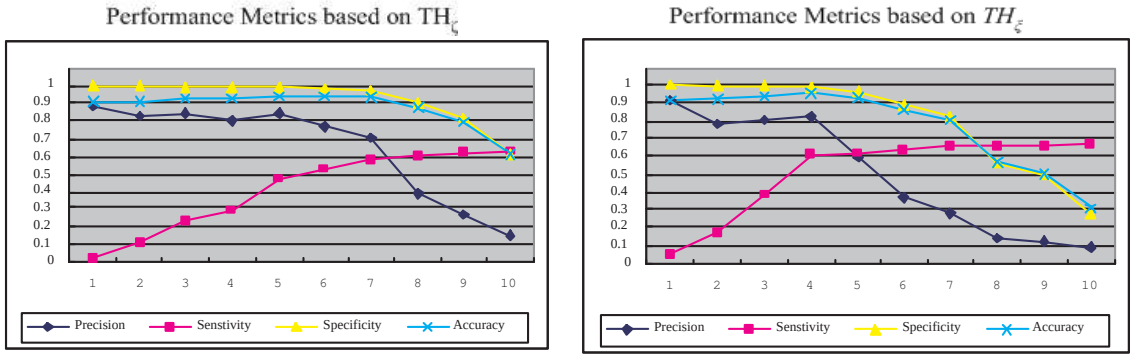

Figure 5.23. (a) Performance Metrics based on Different TH_ζ, (b) Performance Metrics based on Different TH_ξ.

By evaluating filtering by edge distance with several combinations of TH_ζ and TH_ξ, we can obtain the best result (i.e. highest accuracy) when $TH_\zeta = 6$ and $TH_\xi = 3$. Figure 5.24 shows the four performance metrics of filtering by edge distance when we use $TH_\zeta = 6$ and $TH_\xi = 3$.

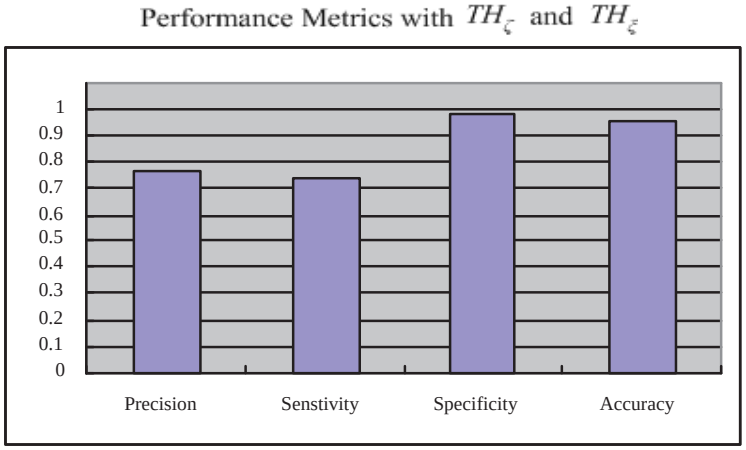

Figure 5.24. Performance Metrics with TH_ζ and TH_ξ.

5.5.3 Evaluation of Polyp Shot Distance

For the polyp shot detection in Section 5.4, we use *Dist* which mainly affects the quality of the polyp shot detection. In this section, we shows how to select the threshold of *Dist* using the same data set of the previous section (Table 5.3).

Table 5.4 shows the four performance metrics of the polyp detection based on different *Dist*. We can obtain the highest accuracy when $Dist = 250$.

5.5.4 Evaluation of Polyp Detection

For this experiment, we extract the three colonoscopy video with 15 frames-per-second rate. We have total 85,520 frames which consists of 6,748 polyp frames and 78,772

Table 5.4. Performance Metrics of Polyp Shot Detection with Different *Dist*

Dist	Precision	Senstivity	Specificity	Accuracy
50	0.774	0.787	0.976	0.958
100	0.771	0.811	0.974	0.959
150	0.776	0.856	0.974	0.963
200	0.766	0.919	0.970	0.965
250	0.764	0.941	0.969	0.967
300	0.664	0.954	0.949	0.950
350	0.474	0.961	0.888	0.895
400	0.388	0.971	0.840	0.853
450	0.326	0.979	0.789	0.807
500	0.260	0.985	0.707	0.734

normal frames. The polyp frame represents a frame containing polyps, and the normal frame represents a frame in which no polyp is included. Table 5.5 shows the detail of our data set.

Table 5.5. Test Set

Video ID	# of Polyp Frames	# of Normal Frames	Total # of Frames
10	3366	22331	25697
102	1186	19731	20917
114	2196	36710	38906
Total	6748	78772	85520

Table 5.6 shows the experimental results of our polyp detection technique. The performance metrics in Row "Without Shot" have been obtained by only our polyp candidate selection technique in Section 5.3, and the performance metrics in Row "With Shot" have been obtained by our polyp candidate selection technique followed by our polyp shot detection in Section 5.4. The performance metrics using the texture-based polyp detection method proposed in [16] are presented in Row "Texture" for comparison.

For the texture-based polyp detection method, we extract Color Covariance features using the 32x32 window size. We use Support Vector machine method in WEKA [42] for the classification.

Even though the precision, specificity and accuracy of "Without Shot" is similar to those of "With Shot", the sensitivity (recall) of "With Shot"(84%) is much higher than the sensitivity of "Without Shot"(74%) because the polyp shot detection method in Section 5.4 can find the missed polyp frames by comparing the initial polyp candidate frames to the adjacent frames. Table 5.6 shows that our proposed technique outperforms the texture-based method.

Table 5.6. Performance Metrics of Polyp Detection

		Precision	Sensitivity	Specificity	Accuracy
010	Without Shot	0.671	0.720	0.975	0.958
	With Shot	0.685	0.803	0.974	0.962
	Texture	0.210	0.732	0.807	0.802
102	Without Shot	0.600	0.835	0.975	0.969
	With Shot	0.601	0.908	0.973	0.970
	Texture	0.210	0.747	0.875	0.870
114	Without Shot	0.630	0.692	0.984	0.973
	With Shot	0.645	0.825	0.982	0.976
	Texture	0.176	0.719	0.868	0.863
Ave	Without Shot	0.634	0.749	0.978	0.967
	With Shot	0.643	0.845	0.976	0.969
	Texture	0.199	0.732	0.850	0.845

Table 5.7 shows the number of polyp shots, the number of correctly-detected polyp shots, the number of falsely-detected polyp shots, and the number of missed polyp shots. Falsely-detected polyp shots represents the shots which do not have any actual polyp but

detected as having polyps by our algorithm. Missed polyp shots are the actual polyp shots but not detected. Among 93 polyp shots, 13 shots are missed and 27 incorrect shots are detected.

Table 5.7. Result of Polyp Shot Detection

ID	Polyp Shots	Correctly-detected	Falsely-detected	Missed
10	35	40	4	9
102	21	23	3	5
114	37	44	6	13
Total	93	107	13	27

We present some examples of polyps which are detected by our algorithm in Figure 5.25. The images at the top are the original polyp images, the images in the middle are the binary edge maps, and the images at the bottom are the detected polyp regions.

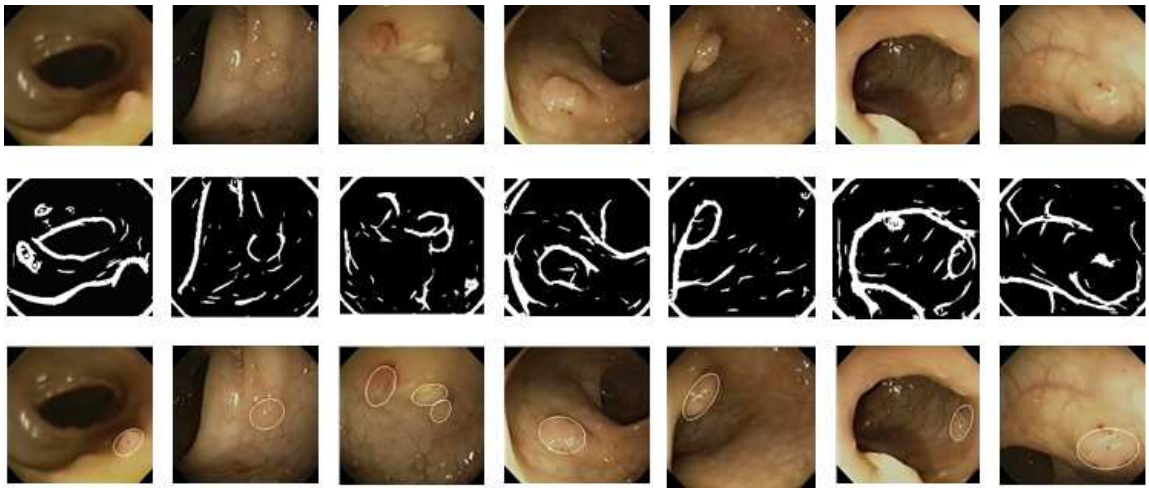

Figure 5.25. Top: Original Polyp Image, Middle: Binary Edge Map, and Bottom: Detected Polyp.

CHAPTER 6

MULTI-LEVEL ENDOSCOPY VIDEO SEGMENTATION

The problem of segmenting visual data into smaller chunks is a basic problem in multimedia analysis, and its solution helps in problems such as video indexing and retrieval. Early video database systems segment video into shots, and extract key frames from each shot to represent it. Such systems have been criticized for not conveying much semantics because they only employ low-level image features to model and index video data, which may cause semantically unrelated data to be close only because they may be similar in terms of their low-level features. More recent approaches construct scene, higher-level abstractions than shots, by grouping a certain number of adjacent shots based on the detected shot pattern. However, these scene segmentation techniques are not suitable for segmenting endoscopy video because endoscopy videos are usually generated by a single camera operation without shot, which makes it difficult to manage and analyze them. In this dissertation, we propose a novel algorithm of multi-level segmentation for endoscopy video, which represents the semantic structure of medical video: Video, Phase, Piece, and Objective shot as depicted in Figure 6.1.

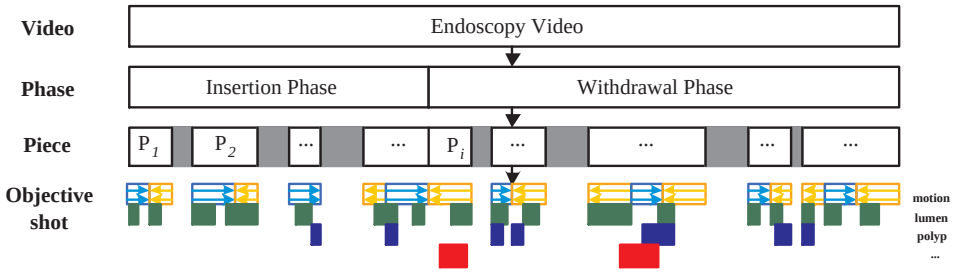

Figure 6.1. Multi-level Segmentation for Endoscopy Video .

Each level of endoscopy video segmentation can be defined as follows:

- Phase: An endoscopic procedure consists of two phases: an insertion phase and a withdrawal phase. During the insertion phase, a flexible endoscope (a flexible tube with a tiny video camera at the tip) is advanced into the most proximal part of the colon or the terminal ileum. In the withdrawal phase, the endoscope is gradually withdrawn. The purpose of the insertion phase is to reach the cecum or the terminal ileum. Careful mucosa inspection and diagnostic or therapeutic interventions are performed in the withdrawal phase.
- Piece: Current endoscopes are equipped with a single, wide-angle lens, and do not have any camera operation function such as zoom-in, zoom-out and auto focusing. Because of these limitations, a significant number of out-of-focus frames (i.e. non-informative frame) are included. By removing non-informative frames, a phase can be decomposed into a number of pieces.
- Objective Shot: A piece can be decomposed into several kinds of shots based on human perception understanding the video contents such as endoscope movement and important objects (i.e. lumen and polyp). Objective shots are constructed by considering the spatio-temporal relationship within a video.

In this chaper, we propose the phase segmentation and the motion shot segmentation technique based on the camera motion estimation as follows. We first detect and discard non-informative frames from the videos, which is presented in Chapter 3. Second, we estimate the camera motions to find a boundary between insertion and withdrawal phases. The insertion phase does not always consist of continuous forward camera motions. The withdrawal phase does not always consist of continuous backward camera motions since the endoscopist constantly moves a camera back and forth to obtain optimal views to inspect the interesting regions such as polyps, cancers, terminal ileum, crowfoot with appendix, ileo-cecal valve, etc. Hence, either phase has an arbitrary num-

ber and combination of forward and backward camera motions while the dominant camera motions of insertion and withdrawal phases are forward and backward, respectively. Third, we segment a colonoscopy video based on the camera motions such as forward and backward, which are called *oral* direction and *anal* direction respectively as described in Figure 6.2. We define a *camera motion shot* as a sequence of consecutive frames with a single direction of camera motion. A camera motion shot can be either an *oral shot* which represents the camera motion from the anus to the terminal ileum (forward camera motion) or an *anal shot* which represents the camera motion from terminal ileum to anus (backward camera motion). By accumulating the values of camera motions in the oral and anal shots in an entire video, and finding a peak value, we can locate the end of insertion phase.

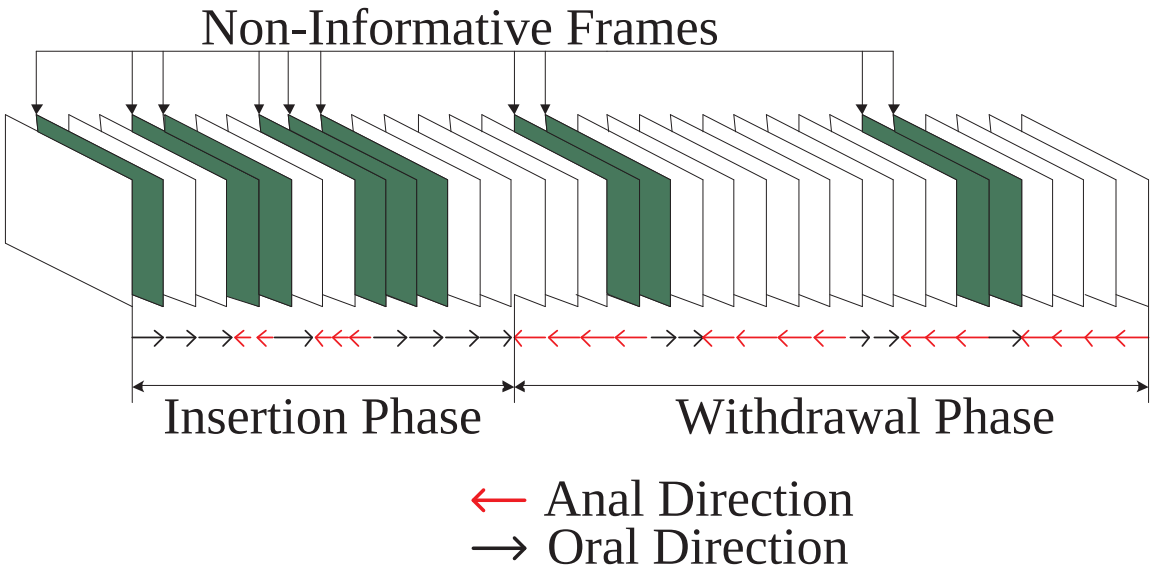

Figure 6.2. Camera Motions in a Colonoscopy Video.

6.1 Camera Motion Estimation

In this section, we present the camera motions (especially, backward camera movement and forward camera movement) based on the affine model in compressed MPEG videos since the provided colonoscopy videos are in MPEG format. After motion vectors are extracted and their outliers are filtered, the camera motions are estimated using the affine model.

6.1.1 Motion Vector Extraction

The motion vectors are extracted directly from the compressed MPEG stream. In MPEG, there are three types of frames: *I-frame*, *P-frame* and *B-frame*. The I-frames are intra coded (i.e. they are encoded without any reference to other frames). The P-frames are encoded based on forward prediction from the previous I-frame or P-frame, and the B-frames are encoded both forward and backward predictions from the previous/next I-frame or P-frame. P-frames and B-frames are referred to as inter coded frames. Only the motion vectors from P-frames are processed in our approach for two reasons. First, usually every third and fifth frame in our MPEG videos is a P-frame, and thus, the temporal resolution is sufficient for our case. Second, both the prediction direction and the temporal distance of motion vectors in B-frames do not exhibit useful patterns for our purposes. Each P-frame consists of a number of macroblocks as shown in Figure 6.3 and each macroblock is associated with a motion vector (mv). The motion vector (mv) which represents the displacement of macroblock between two consecutive frames is extracted from a pair of consecutive frames.These motion vectors are already present in the video because the MPEG videos are encoded using these motion vectors so we need to extract this motion vector for all the macroblocks in the frames. This can save a significant computation.

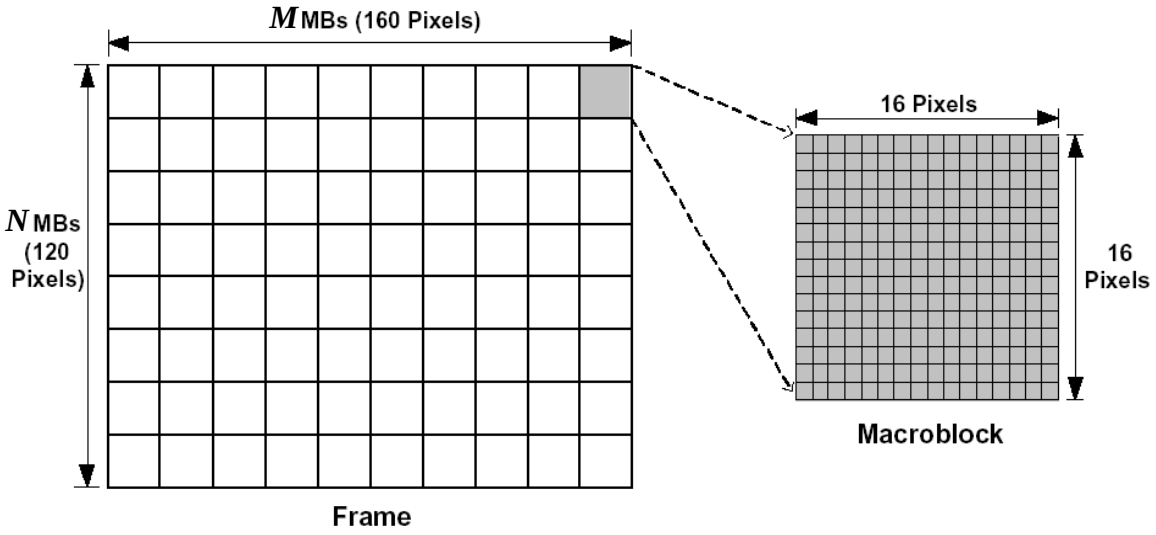

Figure 6.3. A Frame and its Macroblocks.

6.1.2 Motion Vector Filtering

To deal with motion vectors that may not be relevant, various outlier removal algorithms have been proposed. One of them is a heuristic method [63], but it is more useful for special cases. A smoothing filter has been used in [64] to handle the general case, but erroneous outliers remain in the motion vector field. We apply a more reliable method presented in [65] on every macroblock to detect the outlier motion vectors. This method consists of two main steps named *smooth change* and *neighborhood*. A motion vector (mv) is declared as an outlier if both the steps declare it as an outlier (see the examples in Figure 6.4). All detected outliers motion vectors are then removed. The two steps for outlier detection are explained as follows.

- Smooth change: The central mv is compared to each average of four pairs of opposite neighbors. If the distance between the average mv of each pair and the central mv is less than a certain threshold, it is considered as a supporting pair. In Figure 6.4 (a), pairs 1 and 3 are supporting pairs so the number of supporting pairs is 2. If the number of supporting pairs is below a threshold, the central mv

is declared as an outlier. For our experiments, we use a value 3 for the threshold of supporting pairs.

- Neighborhood: A neighborhood motion vector supports the central mv if it lies within a tolerance angle (see Figure 6.4 (b)). If the number of supporting vectors is below a threshold, then the central mv is declared as an outlier. For our experiments, we use a value 4 for the threshold of supporting vectors.

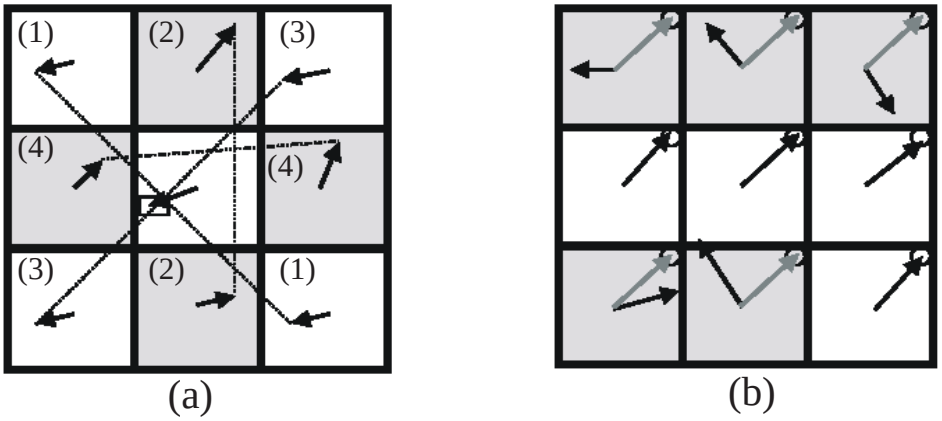

Figure 6.4. Patterns for Motion Vectors Filtering: (a) Smooth Change and (b) Neighborhood.

6.1.3 Camera Motion Estimation

In this section, we discuss the camera motion estimation in compressed MPEG video. A pattern-based approach [66] estimates camera motions based on the analysis of motion vector fields by matching motion vector fields with predefined models in Hough space. Different camera motions will be recognized by comparing the computed results with the prior known patterns. However, such predefined pattern-based approaches are noise sensitive and computationally intensive. In addition, they do not estimate the magnitudes of camera motions. More robust camera motion estimation methods based

on mathematical models such as affine flow, planar surface flow, general optical flow, etc., have been presented [67, 68, 69]. The affine model is known to be more resilient to noise, sparse motion vector fields, and representing all basic types of camera motions. As seen in Figure 6.5, seven camera motions can be defined as follows.

$$\begin{aligned}
Tracking &: translation\ along\ X\ axis;\\
Booming &: translation\ along\ Y\ axis;\\
Dolling &: translation\ along\ Z\ axis;\\
Tilting &: rotation\ along\ X\ axis;\\
Panning &: rotation\ along\ Y\ axis;\\
Rolling &: rotation\ along\ Z\ axis;\\
Zooming &: change\ of\ the\ focal\ length\ (f);
\end{aligned}$$

All seven camera motions can be expressed in the affine model as follows.

$$\begin{aligned}
\begin{pmatrix} u \\ v \end{pmatrix} = & \begin{pmatrix} a_1^{zoom} & b_1^{roll} \\ -b_2^{roll} & a_2^{zoom} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} c^{pan} \\ d^{tilt} \end{pmatrix} \\
& + \frac{1}{z} \left(\begin{pmatrix} a_1^{dolly} & 0 \\ 0 & a_2^{dolly} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} c^{track} \\ d^{boom} \end{pmatrix} \right)
\end{aligned} \tag{6.1}$$

where (u, v) is the motion vector of a macroblock located at position (x, y) of each frame, z is the depth of the real world, a_1^{zoom}, b_1^{roll}, a_2^{zoom}, b_2^{roll}, c^{pan}, d^{tilt}, a_1^{dolly}, a_2^{dolly}, c^{track} and d^{boom} are scalar coefficients concerned with camera motions. Since the endoscopes do not have zoom-in and zoom-out functions, a_1^{zoom}=0 and a_2^{zoom}=0. So Equation (1) can be rewritten as follows.

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{a_1^{dolly}}{z} & b_1^{roll} \\ -b_2^{roll} & \frac{a_2^{dolly}}{z} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} c^{pan} + \frac{c^{track}}{z} \\ d^{tilt} + \frac{d^{boom}}{z} \end{pmatrix} \tag{6.2}$$

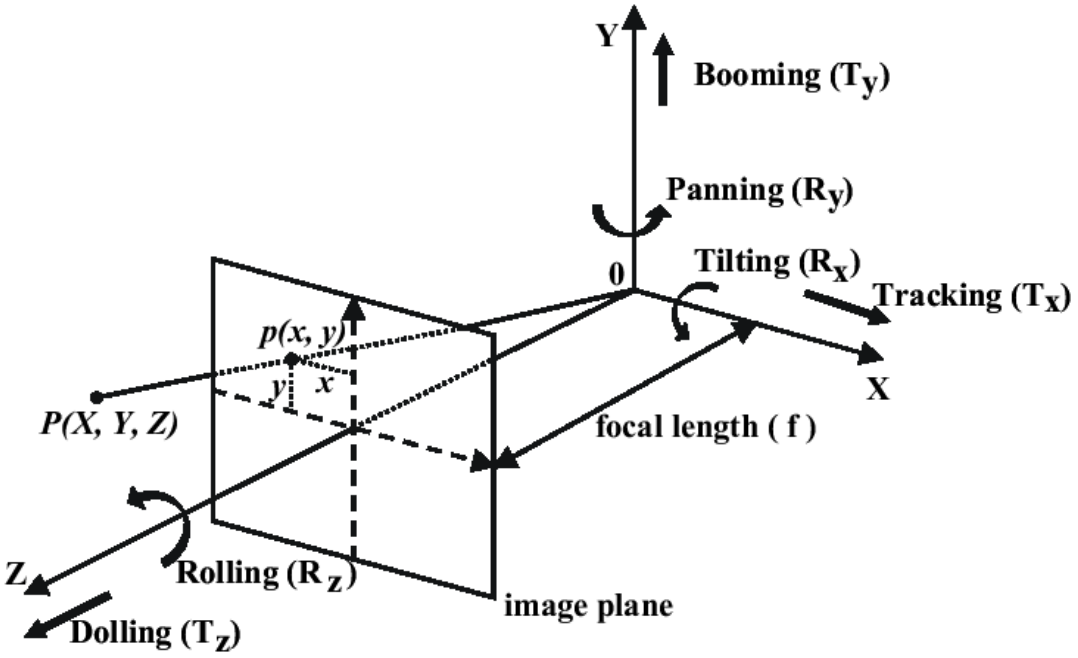

Figure 6.5. 3D Camera Motion Model.

Let

$$a_1 = c^{pan} + \frac{c^{track}}{z}, \quad a_2 = \frac{a_1^{dolly}}{z}, \quad a_3 = b_1^{roll}$$
$$a_4 = d^{tilt} + \frac{d^{boom}}{z}, \quad a_5 = -b_2^{roll}, \quad a_6 = \frac{a_2^{dolly}}{z}.$$

Equation (2) can then be rewritten as follows.

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} a_2 & a_3 \\ a_5 & a_6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_1 \\ a_4 \end{pmatrix} \tag{6.3}$$

Given the motion vectors, we calculate the parameter values $\{a_1, a_2, a_3, a_4, a_5, a_6\}$ using the Least Square Fitting method. Let $\hat{u}$ and $\hat{v}$ be the estimated motion vectors,

then the distance between the estimated motion vector $(\hat{u}, \hat{v})$ and the extracted motion vector (u, v) is

$$\begin{aligned} Dist &= \sum_x \sum_y [(\hat{u}-u)^2 + (\hat{v}-v)^2] \\ &= \sum_x \sum_y [(\hat{u}-(a_1+a_2x+a_3y))^2 + (\hat{v}-(a_4+a_5x+a_6y))^2] \end{aligned}$$

The parameter values are obtained when $Dist$ is minimized and the condition for $Dist$ to be minimized is the first derivative of $Dist$ to 0 such as

$$\frac{\partial Dist}{\partial a_1} = 0, \frac{\partial Dist}{\partial a_2} = 0, \frac{\partial Dist}{\partial a_3} = 0, \frac{\partial Dist}{\partial a_4} = 0, \frac{\partial Dist}{\partial a_5} = 0, \frac{\partial Dist}{\partial a_6} = 0$$

By solving the above equations, we can get the parameter values as follows.

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} N & A & B \\ A & C & E \\ B & E & D \end{pmatrix}^{-1} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix}, \quad \begin{pmatrix} a_4 \\ a_5 \\ a_6 \end{pmatrix} = \begin{pmatrix} N & A & B \\ A & C & E \\ B & E & D \end{pmatrix}^{-1} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix}$$

where,

$$\begin{aligned} N = \sum_x \sum_y 1, \quad A = \sum_x \sum_y x, \quad B = \sum_x \sum_y y \\ C = \sum_x \sum_y x^2, \quad D = \sum_x \sum_y y^2, \quad E = \sum_x \sum_y xy \\ U_1 = \sum_x \sum_y u, \quad U_2 = \sum_x \sum_y ux, \quad U_3 = \sum_x \sum_y uy \\ V_1 = \sum_x \sum_y v, \quad V_2 = \sum_x \sum_y vx, \quad V_3 = \sum_x \sum_y vy \end{aligned}$$

Even though the parameter values $\{a_1, a_2, a_3, a_4, a_5, a_6\}$ are related with camera motions, we can obtain more reliable camera motions such as Dolling Camera Motion (DCM), Rolling Camera Motion (RCM), Horizontal Camera Motion (HCM=Panning+Tracking), and Vertical Camera Motion (VCM=Tilting+Booming) as follows.

$$DCM = \frac{1}{2}(a_2 + a_6), \quad HCM = a_1, \quad RCM = \frac{1}{2}(a_3 - a_5), \quad VCM = a_4$$

Among the four camera motions, the dolling camera motion (DCM) will mainly be examined because the dolling camera motion is directly related to the forward and backward camera movement. The positive DCM value means forward movement and the negative DCM value means backward movement.

6.2 Phase and Motion Shot Segmentation

Using the dolling camera motion (DCM), we can easily segment a colonoscopy video into a number of shots, each of which consists of the frames with the same camera motion (oral shot or anal shot). However, the phase segmentation needs more steps because all frames in the insertion phase are not oral direction frames, and all frames in the withdrawal phase are not anal direction frames. Even though, all frames do not have the same directional camera movements in each phase, the insertion phase consists of a large number of oral shots (oral direction frames) and a small number of anal shots (anal direction frames), and the withdrawal phase consists of a large number of anal shots (anal direction frames) and a small number of oral shots (oral direction frames). Using these characteristics of colonoscopy videos, we propose the following video segmentation method.

1. **Non-informative frame filtration step** removes non-informative frames.
2. **Camera motion estimation step** calculates four camera motions (Dolling Camera Motion (DCM), Rolling Camera Motion (RCM), Horizontal Camera Motion (HCM), and Vertical Camera Motion (VCM)) for the informative frames.
3. **Unreliable DCM value filtration step** filters out unreliable DCM values as follows.
    - If there are few motion vectors (mv) between two consecutive frames, an abrupt change exists between them and the estimated camera motions are not correct. We remove this type of errors by assigning $DCM = 0$ if the number

of motion vectors is less than a certain threshold value ($mv < TH_\eta$). In our experiments, we use a value 10 for TH_η.

- The DCM tends to have an incorrect value when other camera motions such as Horizontal Camera Motion (HCM) or Vertical Camera Motion (VCM) have bigger values compared with DCM. To reduce this type of errors, we assign $DCM = 0$ if the ratio of the magnitudes of HCM and VCM to DCM ($\frac{\sqrt{HCM^2+VCM^2}}{DCM}$) is larger than a certain threshold (TH_ζ). In our experiments, we use a value 1500 for TH_ζ.
- Temporal information is utilized to filter out incorrect DCMs. It is highly likely that any oral or anal shots have more than two frames (we are using 30 frames/second rate videos). Therefore, we assign $DCM = 0$ if the number of consecutive frames with the same direction is less than a certain threshold (TH_δ). In our experiments, we use a value 2 for TH_δ.

4. **Shot boundary detection step** detects shot boundaries of a colonoscopy video based on camera movements. As seen in Figure 6.6 (a), a colonoscopy video can be decomposed into a number of pieces ($P_1, P_2, \cdots, P_i, \cdots$) by the non-informative frame filtration. Each piece consists of a number of frames with three different kinds of DCM values: frames with positive DCM values, frames with negative DCM values and frames with DCM=0. Using the DCM values of frames in a piece (P_i), we detect shot boundaries as follows.
   (a) Let P_i have n numbers of frames ($F_1^i, F_2^i, \cdots, F_n^i$) and let the DCM values of these frames be DCM_1^i,

   $DCM_2^i, \cdots, DCM_n^i$. We consider two frames at a time: F_p^i and F_q^i. Initially, we set p=1 and q=2.

(b) Check if the DCM value of F_p^i is zero ($DCM_p^i = 0$). If $DCM_p^i = 0$, increment p and q by 1 ($p = p+1$ and $q = q+1$) until the DCM value of F_p^i is not zero (i.e., forward movement or backward movement exists.)

(c) Compare the DCM value of F_p^i (DCM_p^i) with the DCM value of F_q^i (DCM_q^i) until F_q^i is the last frame of P_i ($q = n$) as follows.

- If $DCM_p^i \times DCM_q^i > 0$, increment p and q by 1 ($p = p+1$ and $q = q+1$).
- If $DCM_q^i = 0$, increase q by 1 ($q = q+1$).
- If $DCM_p^i \times DCM_q^i < 0$, a shot boundary is detected between F_{q-1}^i and F_q^i. Two frames (F_p^i and F_q^i) are reset such as $p = q$ and $q = q+1$.

Figure 6.6 (b) shows an example of the detected shot boundaries using the above process.

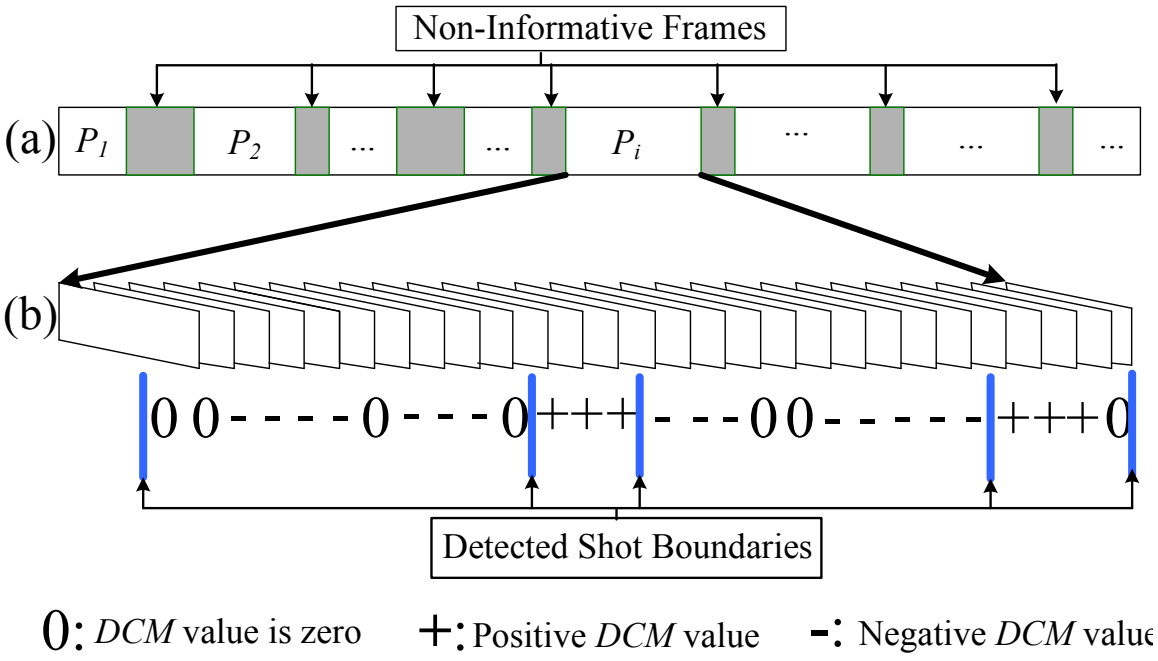

Figure 6.6. Example of Shot Boundary Detection.

5. **Phase boundary detection step** detects the boundary between the insertion phase and the withdrawal phase using the accumulated DCM. When we add up all DCM values, the accumulated DCM value will increase until the end of the insertion phase because most of the frames in the insertion phase have forward movements (i.e., positive DCM values). However, the accumulated DCM value will decrease during the withdrawal phase because most of the frames in the withdrawal phase have the backward movement (i.e., negative DCM values). For this reason, the boundary frame between the insertion phase and the withdrawal phase has the highest accumulated DCM values.

Figure 6.7 shows an example of video segmentation obtained using our shot and phase segmentation method.

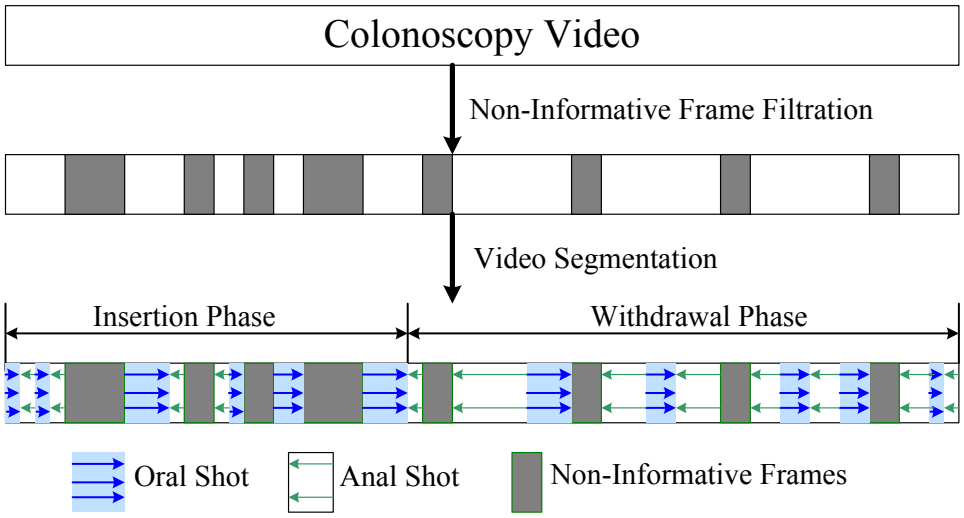

Figure 6.7. Example of Video Segmentation.

6.3 Experimental Results

In this section, we report our experimental results using videos we produced in a control environment and real colonoscopy videos. The produced videos are used for evaluation of the performance of the camera motion estimation since the ground truth of camera motions in these videos are easy to determine. The data set of colonoscopy videos is used to evaluate the quality metrics.

6.3.1 Evaluation of Camera Motion Estimation

The effectiveness of our video segmentation and related quality metrics depends on the accuracy of the estimated camera motions. To evaluate the camera motion estimation technique, we produced six videos tracing the same distance of a long corridor without zoom-in and zoom-out camera motions, which resemble the contents of colonoscopy video. Each of them has different content(s) based on camera motion, and the details are shown in Table 6.1.

Table 6.1. Test Set I: Produced Videos

ID	Contents	# of Frames
Video1	Fast Forward	275
Video2	Slow Forward	617
Video3	Fast Backward	265
Video4	Slow Backward	548
Video5	Forward-Backward-Forward	833
Video6	Backward-Forward-Backward	788

The column "ID" in Table 6.1 represents the unique ID for each video, and the column "Contents" represents the content of each video. For example, "Fast Forward" means that the video was recorded with only forward camera movement with relatively high speed, and "Slow Backward" means that the video was recorded with only back-

ward camera movement with relatively low speed. "Forward-Backward-Forward" and "Backward-Forward-Backward" indicate the order of camera movements in each video. Figure 6.8 shows some examples of Video1. We note that 30 frames are extracted per second in our experiments.

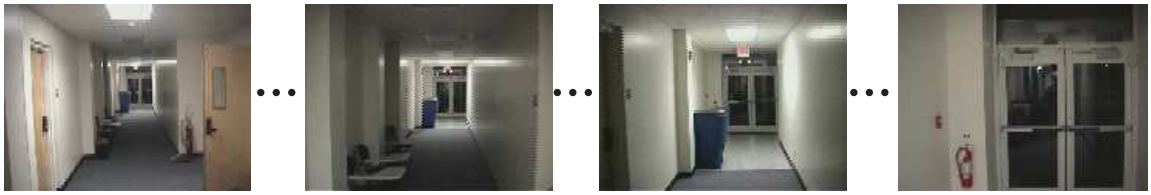

Figure 6.8. Examples of Fast Forward Camera Movement.

Table 6.2 shows the error rate of the detected camera motion. The error rate (ER) can be estimated by the number of frames in which their motions are incorrectly detected divided by the total number of frames as follows.

$$ER = \frac{\text{Number of Incorrect Motion Frames}}{\text{Total Frames}} * 100(\%)$$

Table 6.2. Error Rate of Camera Motion Detection

ID	ERNOR	EROR
Video1	5.80	1.45
Video2	10.46	0.65
Video3	18.46	1.54
Video4	14.60	0.73
Video5	8.65	1.44
Video6	9.64	1.02
Total	10.74	1.09

In Table 6.2, the label "NOR" means no outlier removal whereas "OR" means outlier removal. The column "ER-NOR" in Table 6.2 represents the error rate of the detected camera motions without outlier removal. The column "ER-OR" represents the error rate of the detected camera motions with outlier removal. The results show that our outlier removal method significantly increases the performance of the camera motion detection by reducing the average error rate (ER) from 10.74% to 1.09%.

We can also verify the accuracy of the detected camera motion by accumulating the Dolling Camera Motion (DCM) values. Figure 6.9 shows the plot of the accumulated DCM values of six different videos. All six videos were recorded with the same physical distance of moving so the last frames of six videos have very similar magnitude DCM values around 1.7 regardless of the speed of the movement in each video. Also, we can detect the boundaries between the forward camera movement, and the backward camera movement by calculating the local maxima of the accumulated DCM values, and the boundaries between the backward camera movement and the forward camera movement by calculating the local minima, respectively. The arrows in the two images at the bottom of Figure 6.9 indicate the local maxima and local minima where the direction of the camera movements changed.

6.3.2 Evaluation of Phase and Motion Shot Segmentation

In this section, we evaluate the effectiveness of our quality metrics with real colonoscopy videos. The total videos in our test set last about 2 hours and 18 minutes, and consist of 249,759 frames (30 frames/sec.). The details are shown in Table 6.3. The column "ID" represents the unique id number for each colonoscopy video, and the column "Total Frames" represents the total number of frames for each video.

First, we report the effectiveness of our non-informative frame filtration and the experiment results are presented in Table 6.4. Column "S" represents the number of ac-

Table 6.3. Test Set II: Colonoscopy Videos

ID	Duration (minute)	Total Frames
100	20	36434
110	15	27659
117	18	33014
148	24	43436
170	20	35599
175	19	33883
180	22	39734

tual non-informative frames we manually determined; column "T" represents the number of detected non-informative frames; and column "C" represents the number of correctly detected non-informative frames. The average precision and recall are 0.953 and 0.958, respectively.

Table 6.4. Effectiveness of Non-Informative Frame Filtration

ID	S	T	C	Precision ($\frac{C}{T}$)	Recall ($\frac{C}{S}$)
100	23099	23392	22267	0.952	0.964
110	12719	12542	12146	0.968	0.955
117	18587	18262	17562	0.962	0.945
148	23063	23166	22091	0.953	0.957
170	18894	19167	18183	0.948	0.962
175	15612	15885	15110	0.951	0.967
180	17834	18040	17024	0.943	0.954
Ave	-	-	-	0.953	0.958

Table 6.5 shows the performance of the shot segmentation technique based on camera motions. Column "S" represents the number of actual shot boundaries we manually determined, column "T" represents the number of detected shot boundaries and column "C" represents the number of correctly detected shot boundaries. Table 6.5 shows the precisions and recalls for our shot segmentation techniques, which are very promising.

Table 6.5. Effectiveness of Shot Detection

ID	S	T	C	Precision($\frac{C}{T}$)	Recall($\frac{C}{S}$)
100	536	574	461	0.803	0.860
110	535	558	470	0.842	0.878
117	434	444	384	0.864	0.884
148	427	487	383	0.786	0.897
170	559	609	501	0.823	0.896
175	395	443	332	0.749	0.841
180	517	557	441	0.792	0.853
Average	-	-	-	0.809	0.873

Using our video segmentation technique, we can detect the boundary between the end of the insertion phase and the beginning of the withdrawal phase, which in the vast majority of cases is characterized by the presence of terminal ileum, crowfoot with appendix or ileo-cecal valve. However, it is not easy to verify whether the detected frame is correct or not without the overview around the frame with highest accumulated DCM value. For this reason, we evaluate the phase detection algorithm using the following criterion:

- *Criterion of Phase Detection:* the phase detection correctly detect the phase boundary if there is the end of a colon (i.e. terminal ileum, crowfoot with appendix or ileo-cecal valve) 30 seconds before and after the detected boundary frame.

We tested the phase segmentation algorithm based on the above criterion using seven colonoscopy videos in Table 6.3, and our phase segmentation algorithm have satisfied the criterion with seven colonoscopy videos. Figure 6.10 shows the accumulated DCM plots of three test videos depicting how our phase segmentation technique correctly finds the boundary between the end of insertion phase and the start of the withdrawal phase utilizing the accumulated DCM. Three colon images are presented below the accumulated DCM plot. The first image is the frame with the highest accumulated DCM value, and

the second and third images are images of frames obtained 1000 or 2000 frames after the first frame, respectively. The frame with the highest accumulated DCM value shows the very proximal of the colon for each colonoscopy video.

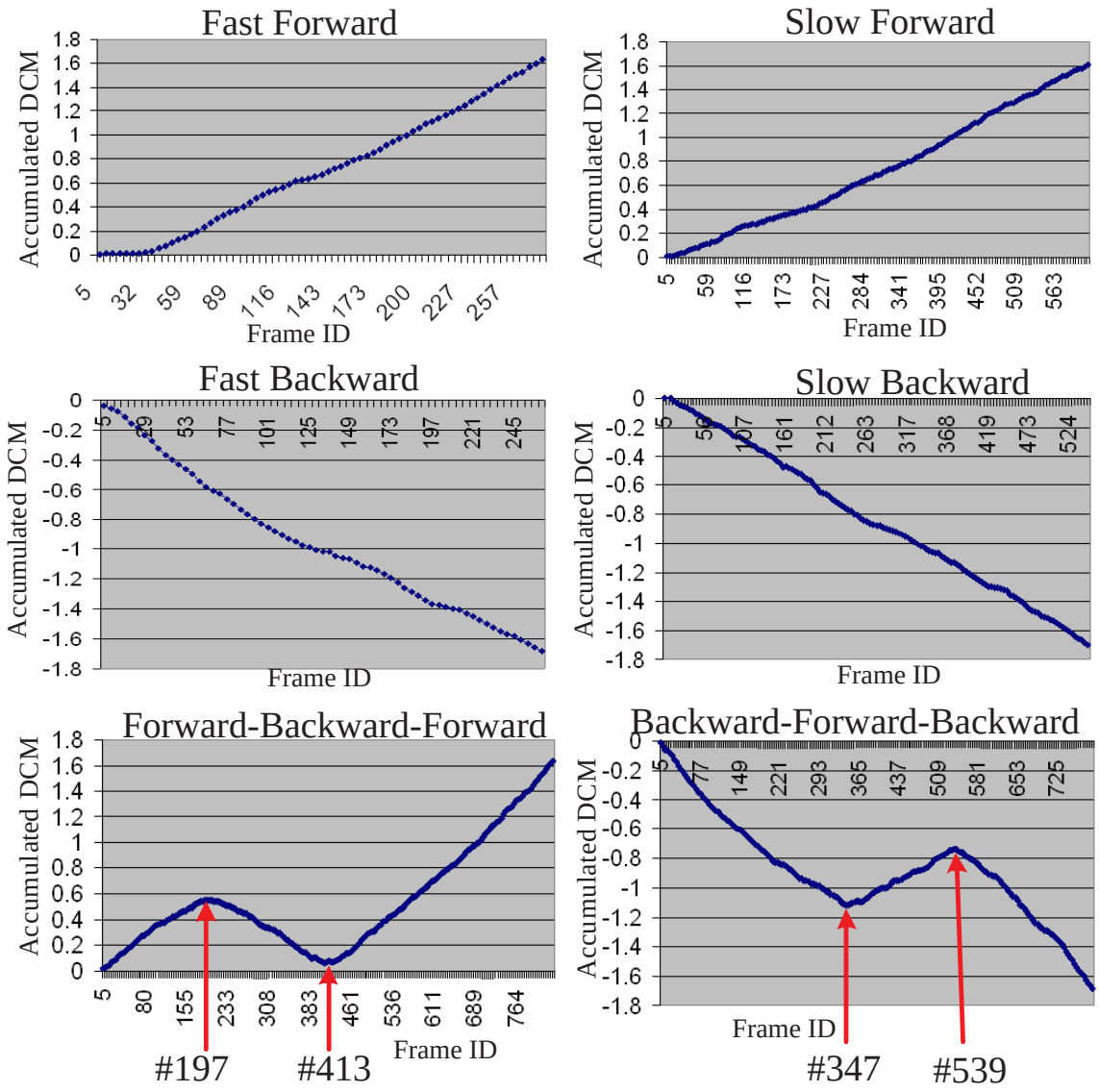

Figure 6.9. Effectiveness of Accumulated DCM.

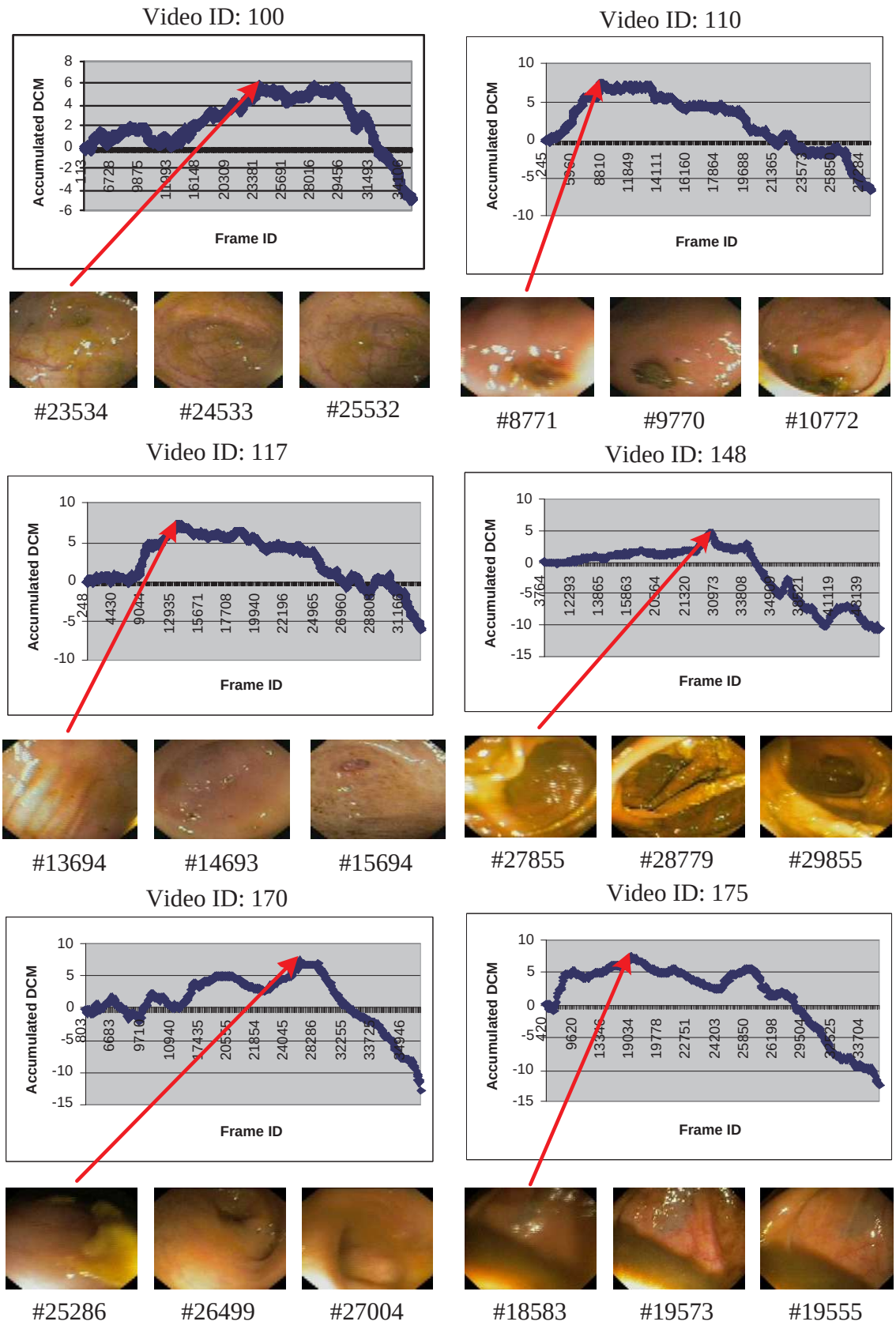

Figure 6.10. Phase Segmentation of Colonoscopy Videos.

CHAPTER 7

MEASUREMENT OF ENDOSCOPY QUALITY

Although colonoscopy has become the preferred screening modality for prevention of colorectal cancer, recent data suggest that there is a significant miss-rate for the detection of even large polyps and cancers [5, 6, 7]. The miss-rate varies among endoscopists and it may be related to the experience of the endoscopist and the location of the lesion in the colon. Even though the demand for quality control in colonoscopic procedures has been gaining force, there is no study related to this have been done thus far. In general, the quality of a colonoscopic procedure can be evaluated in terms of time of the withdrawal phase and thoroughness of inspection of the colon mucosa. Current American Society for Gastrointestinal Endoscopy guidelines suggest that on average the withdrawal phase during a screening colonoscopy should last a minimum of 6-10 minutes. The main purpose of this chapter is to develop new objective metrics from automatic analysis of a colonoscopy video to evaluate the endoscopist's skill and the quality of colonoscopy. Our invention on this automatic quality measurement system has been selected by American College of Gastroenterology (ACG) for ACG Governors Award for Excellence in Clinical Research in 2006.

7.1 Quality Metrics

In this section, we investigate the five objective metrics to evaluate the endoscopist's skill and the quality of colonoscopy. The information to calculate the quality metrics are obtained from our multi-level endoscopy video segmentation techniques as depicted in

Figure 7.1. We note that the superscript of each metric represents the phase ID (i.e., O for insertion phase and A for withdrawal phase)

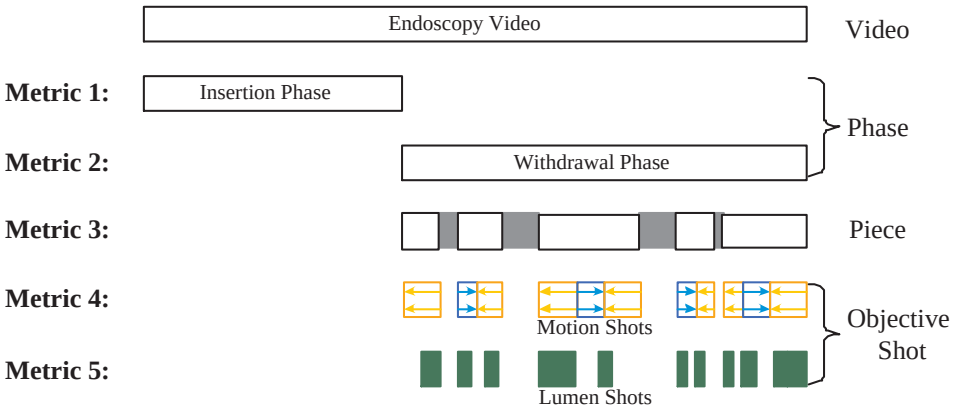

Figure 7.1. Overview of Quality Measure .

- **Metric 1:** The purpose of the insertion phase is to reach the proximal end of the colon. The insertion time (IT) can be measured as follows:

$$IT = \frac{\text{NF}^O}{\text{Frame Extraction Rate}}$$

where NF^O represents the number of frames in the insertion phase.

- **Metric 2:** The withdrawal time (WT) is the duration of the withdrawal phase. We calculate WT as follows.

$$WT = \frac{\text{NF}^A}{\text{Frame Extraction Rate}}$$

where NF^A represents the number of frames in the withdrawal phase.

- **Metric 3:** Even though the duration of the withdrawal phase is long, we cannot say that the quality of the colonoscopy is good if a colonoscopy has a lot of non-informative frames in the withdrawal phase. By adding up the duration of only the

informative frames in the withdrawal phase, we can obtain the clear withdrawal time (CWT) and the clear withdrawal ratio (CWR) computed as follows:

$$CWT = \frac{\text{NIF}^A}{\text{Frame Extraction Rate}}, \quad CWR = \frac{CWT}{WT}$$

where NIF^A represents the number of informative frames in the withdrawal phase.

- **Metric 4:** During the withdrawal phase, an endoscopist may move a camera back and forth to examine suspicious regions. This movement may be an indicator of quality, as the endoscopist is trying to verify that an area is indeed free of lesions on multiple inspections. These movements can be estimated by measuring the number of camera motion changes in the withdrawal phase ($NCMC$), and the ratio of the number of camera motion changes to the clear withdrawal time ($RCMC$) as follows.

$$NCMC = \text{NS}^A, \quad RCMC = \frac{NCMC}{CWT}$$

where NS^A represents the number of oral and anal shots in the withdrawal phase.

- **Metric 5:** We measure the wall-lumen inspection ratio ($WLIR$) to see whether the endoscopist has appropriate ratio of both close inspections and global inspections in which the colon lumen is seen. We also measure the wall inspection fraction (WIF) to reveal the fraction of the clear withdrawal time spent on examining the colon walls very closely.

$$WLIR = \frac{\text{NWV}^A}{\text{NLV}^A}$$
$$WIF = \frac{\text{NWV}^A}{\text{NIF}^A}$$

where NWV^A represents the number of the wall view frames in the withdrawal phase and NLV^A represents the number of the lumen view frames in the withdrawal phase.

7.2 Quality Metric Report

We calculate the five metrics for the seven videos using the information obtained from the multi-level endoscopy video segmentation (Table 7.1). Column "NF" represents the number of frames; column "NIF" represents the number of informative frames; and column "NS" represents the number of shots. The superscript of each column represents the phase ID (i.e., O for insertion phase and A for withdrawal phase).

Table 7.1. Information of Colonoscopy Videos

ID	Insertion Phase	Withdrawal Phase				
	NF^O	NF^A	NIF^A	NS^A	NWV^A	NLV^A
100	23534	12900	7427	147	1853	5574
110	8771	18888	12836	262	4147	8689
117	13694	19320	11748	186	3467	8281
148	27855	15581	9130	219	2465	6665
170	25286	7676	7114	202	1594	5520
175	18583	15300	9363	207	2277	7086
180	27043	12691	8477	201	2219	6258

The generated quality metrics from automatic analysis of videos in Table 7.1 are shown in Table 7.2.

- Metric 1: We compute the insertion time $IT = \mathrm{NF}^O/30$ because we extract 30 frames per second for our experiments. The results are summarized in the second column of Table 7.2. The colonoscopy video 110 has a short insertion time (about 5 minutes) and colonoscopy video 148 has a long insertion time (about 15 minutes). Many foreign substances such as stool were found during the insertion phase of colonoscopy video 148 so it was more difficult for the endoscopist to reach the proximal end of the colon. Therefore, it has a bigger IT value.

- Metric 2: The withdrawal time (WT) is computed as $WT = \mathrm{NF}^A/30$. Current American Society for Gastrointestinal Endoscopy guideline suggests that on average the withdrawal time should last a minimum of 6-10 minutes. The colonoscopy video 170 has slightly shorter withdrawal time (i.e. 5 minutes 50 seconds) than this guideline. The details are listed in the third column of Table 7.2.
- Metric 3: We measure the clear withdrawal time as $CWT = \mathrm{NIF}^A/30$ and the ratio of the clear withdrawal time to the withdrawal time ($CWR = CWT/WT$). The colonoscopy video 170 also has shortest clear withdrawal time (i.e. 3 minutes 57 seconds) among seven colonoscopy videos. However, the colonoscopy video 170 consists of more clear frames than other videos because it has the highest ratio of the clear withdrawal time to the entire withdrawal time (CWR=0.692). The withdrawal time of colonoscopy video 117 is a little bit longer than that of colonoscopy video 110, but the clear withdrawal time of colonoscopy video 110 is longer than that of colonoscopy video 117. The details are found in the forth and fifth columns of Table 7.2.
- Metric 4: We measure the number of the camera motion changes ($NCMC = \mathrm{NS}^A$) and the ratio of the number of the camera motion changes to the clear withdrawal time ($RCMC = NCMC/CWT$). There are some regions in colonoscopy video 110 and 180 that the endoscopist apparently can not see well so the endoscopist frequently moves a camera back and forth to examine these regions in order to get the best possible view. Colonoscopy video 180 has bigger values of $NCMC$ and $RCMC$ than the other videos so we can expect that the colonoscopy video 180 represents a colon that is different from the other two colons, and may contain more angulations or haustrae which require more efforts in order to achieve optimal

mucosal inspection. The details are presented in the sixth and seventh columns of Table 7.2.

- Metric 5: We measure wall-lumen inspection ratio ($WLIR = \text{NWV}^A/\text{NLV}^A$) and the wall inspection fraction ($WIF = \text{NWL}^A/\text{NIF}^A$). Colonoscopy video 110 has bigger values of $WLIR$ and WIF than the other videos so we can expect that the colonoscopy video 110 shows more colon mucosa frames than the other videos. The details are presented in the eighth and ninth columns of Table 7.2.

Table 7.2. Automated Quality Metrics

ID	*IT* (min:sec)	*WT* (min:sec)	*CWT* (min:sec)	*CWR*	*NCMC*	*RCMC*	*WLIR*	*WIF*
100	13:4	7:10	4:7	0.576	147	0.594	0.332	0.250
110	4:52	10:29	7:7	0.680	262	0.612	0.477	0.323
117	7:36	10:44	6:31	0.608	186	0.475	0.419	0.295
148	15:28	8:39	5:4	0.586	219	0.422	0.370	0.270
170	14:2	5:50	3:57	0.692	202	0.577	0.289	0.224
175	10:19	8:30	5:12	0.612	207	0.406	0.321	0.243
180	15:1	7:3	4:42	0.668	201	0.686	0.355	0.262

REFERENCES

[1] S. Kuwada, "Colorectal Cancer 2000," *Postgraduate Medicine*, vol. 107, pp. 96–107, March 2000.

[2] S. Phee and W. Ng, "Automatic of colonoscopy: Visual control aspects," *Medicine and Biology Magazine*, 1998.

[3] C. K. Koh and D. F. Gillies, "Using Fourier information for the detection of the lumen in endosocopy images," in *TECON' 94*, 1994, pp. 981–985.

[4] P. Dario and M. C. Lencioni, "A microrobotic system for colonoscopy," in *Proc. of the IEEE international Conference on Robotic and Automation*, Florence, Italy, 1997, pp. 1567–1572.

[5] D. L. Editorial, "Quality and colonoscopy: a new imperative," *Gastrointestinal endoscopy*, vol. 61, no. 2, 2005.

[6] A. Pabby, R. E. Schoen, J. L. Weissfeld, R. Burt, J. W. Kikendall, P. Lance, E. Lanza, and A. Schatzkin, "Analysis of colorectal cancer occurence during surveillance colonoscopy in the dietary prevention trial," *Gastrointestinal Endoscopy*, vol. 61, no. 3, pp. 385–391, 2005.

[7] D. K. Rex, J. H. Bond, S. Winawer, T. R. Levin, R. W. Burt, and D. A. J. et al, "Quality in the technical performance of colonoscopy and the continuous quality improvement process for colonoscopy: recommendations of the u.s. multi-society task force on colorectal cancer," *American Journal Gastroenterol*, vol. 97, no. 6, pp. 1296–1308, 2002.

[8] G. Khan, "A highly parallel shade image segmentation method," in *International Conference on Parallel Processing for Computer Vision and Display*, 1988.

[9] S. Kumar, K. V. Asari, and D. Radhakrishnan, "Online Extraction of Lumen Region and Boundary from Endoscopic Images Using a Quad Structure," in *Proc. of Conf. on Image Processing and Application*, 1999, pp. 818–822.

[10] H. Tian, T. Srikanthan, and K. V. Asari, "Automatic Segmentation Algorithm for the Extraction of Lumen Region and Boundary from Endoscopic Images," *Medical and Biological Engineering and Computing*, vol. 39, pp. 8–14, 2001.

[11] L. E. Sucar and D. F. Gillies, "Knowledge-based assistant for colonoscopy," in *Proc. of the 3rd Int'l Conf. on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, June 1990.

[12] J. Shuttleworth, A. Todman, R. Naguib, B. Newman, and M. Bennett, "Colour texture analysis using co-occurrence matrices for classification of colon cancer images," in *Electrical and Computer Engineering, 2002*, May 2002, pp. 1134–1139.

[13] N. Esgiar, R. N. G. Naguib, B. S. Sharif, M. K. Bennett, and A. Murray, "Microscopic Image Analysis for Quantitative Measurement and Feature Identification of Normal and Cancerous Colonic Mucosa," *IEEE Transactions on Information Technology in Biomedicine*, vol. 2, pp. 197–203, 1998.

[14] S. Karkanis, D. Iakovidis, D. Maroulis, D. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE Transactions on Information Technology in Biomedicine*, vol. 7, no. 141-152, September 2003.

[15] P. M. Tjoa and M. S. Krishnan, "Feature extraction for the analysis of colon status from the endoscopic images," *BioMedicalEngineering OnLine*, April 2003.

[16] D. K. Iakovidis, D. E. Maroulis, S. A. Karkanis, and A. Brokos, "A Comparative Study of Texture Features for the Discrimination of Gastric Polyps in Endoscopic Video," in *Proc. of the 18th IEEE Symposium on Computer-Based Medical Systems*, Ireland, June 2005, pp. 575–580.

[17] S. Lakare, D. Chen, L. Li, A. Kaufman, and J. Liang, "Electronic Colon Cleansing Using Segmentation Rays for Virtual Colonoscopy," in *Proc. of SPIE 2002 Symposium on Medical Imaging*, San Diego, CA, February 2002.

[18] L. Li, D. Chen, S. Lakare, K. Kreeger, I. Bitter, A. Kaufman, M. R. Wax, P. M. Djuric, and Z. Liang, "An Image Segmentation Approach to Extract Colon Lumen through Colonic Material Tagging and Hidden Markov Random Field Model for Virtual Colonoscopy," in *Proc. of SPIE 2002 Symposium on Medical Imaging*, San Diego, CA, February 2002.

[19] S. B. Gokturk, C. Tomasi, B. Acar, C. F. Beaulieu, D. S. Paik, R. B. J. Jr., J. Yee, and S. Napel, "A statistical 3-D pattern processing method for computer-aided detection of polyps in CT colonography," *IEEE Transactions on Medical Imaging*, vol. 20, pp. 1251–1260, December 2001.

[20] R. M. Summers, C. F. Beaulieu, L. M. Pusanik, J. D. Malley, R. B. Jeffrey, D. I. Glazer, and S. Napel, "Automated polyp detector for CT colonography: Feasibility study," *Radiology*, vol. 216, no. 1, pp. 284–290, 2000.

[21] S. Banerjee and J. Dam, "Colonoscopy or CT colonography for colorectal cancer screening in 2006," *Nature Clinical Practice Gastroenterology and Hepatology*, vol. 3, pp. 296–297, 2006.

[22] D. Kundur and D. Hatzinakos, "Blind Image Deconvolution," *IEEE Signal Processing Magazine*, vol. 13, no. 3, pp. 43–64, May 1996.

[23] G. Ayers and J. Dainty, "Iterative blind deconvolution method and its applications," *Optics Letters*, vol. 13, no. 7, pp. 547–549, July 1988.

[24] B. McCallum, "Blind deconvolution by simulated annealing," *Optics Communication*, vol. 75, no. 2, pp. 101–105, February 1990.

[25] R. Bates and H. Jiang, "deconvolution -recovering the seemingly irrecoverable!" *International Trends in Optics*, pp. 423–437, 1991.

[26] R. Nakagaki and A. Katsaggelos, "A VQ-based blind image restoration algorithm," *IEEE Transactions on Image Processing*, vol. 12, no. 9, pp. 1044 – 1053, September 2003.

[27] H. Pai and A. Bovik, "On eigenstructure-based direct multichannel blind image restoration," *IEEE Transactions on Image Processing*, pp. 1434–1446, October 2001.

[28] G. Giannakis and R. J. Heath, "Blind identification of multichannel FIR blurs and perfect image restoration," *IEEE Transactions on Image Processing*, vol. 9, no. 11, pp. 1877 – 1896, November 2000.

[29] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, November 1986.

[30] R. W. Ramirez, *The FFT, fundamentals and concepts.* Prentice-Hall, 1985.

[31] J. S. Walker, *Fast Fourier transforms.* CRC Press, 1996.

[32] M. A. Sid-Ahmed, *Image processing: theory, algorithms, and architectures.* New York, NY, 1995.

[33] R. C. Gonzalez, *Digital image processing.* Prentice Hall, 2002.

[34] M. Sonka, *Image processing, analysis, and machine vision.* PWS Pub, 1999.

[35] R. Haralick, K. Shanmugam, and D. I., "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, pp. 610–621, 1973.

[36] M. Bevk and I. Kononenko, "A statistical approach to texture description of medical images: a preliminary study," in *Proceedings of IEEE Symposium on Computer-Based Medical Systems*, June 2002, pp. 239–244.

[37] J. Felipe, A. Traina, and C. J. Traina, "Retrieval by content of medical images using texture for tissue identification," in *Proceedings of IEEE Symposium on Computer-Based Medical Systems*, June 2003, pp. 175–180.

[38] J. S. Weszka, C. Dyer, and A. Rosenfeld, "A Comparative Study of Texture Measures for Terrain Classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 6, no. 4, pp. 269–285, April 1976.

[39] R. W. Conners and C. A. Harlow., "A theoretical comparison of texture algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 3, pp. 204–222, 1980.

[40] M. Hall-Beyer, "GLCM Texture: A Tutorial," *National Council on Geographic Information and Analysis Remote Sensing Core Curriculum*, 2000.

[41] J. Han and M. Kamber, *Data mining: concepts and techniques.* Morgan Kaufmann Publishers, 2001.

[42] I. H. Witten and L. Frank, *Data mining: practical machine learning tools and techniques with Java implementations.* Morgan Kaufmann Publishers, 2000.

[43] J. van Zyl and I. Cloete, "The influence of the number of clusters on randomly expanded data sets," in *International Conference on Machine Learning and Cybernetics 2003*, November 2003, pp. 355 – 359.

[44] H. Xu and M. Liao, "Cluster-based texture matching for image retrieval," in *International Conference on Image Processing*, 1998, pp. 766 – 769.

[45] C. Chen, J. Luo, and K. Parker, "Image segmentation via adaptive K-means clustering and knowledge-based morphological operations with biomedical applications," *IEEE Transactions on Image Processing*, pp. 1673–1683, 1998.

[46] T. Bhangale, U. Desai, and U. Sharma, "An unsupervised scheme for detection of microcalcifications on mammograms," in *International Conference on Image Processing*, 2000, pp. 184 – 187.

[47] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms (2nd ed).* MIT Press, 2001.

[48] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2001.

[49] R. Halir and J. Flusser, "Numerically Stable Direct Least Squares Fitting of Ellipses," in *Int. Confonference in Central Europe on Computer Graphics, Visualization and Interactive Digital Media*, 1998, pp. 125–132.

[50] W. Gander, "Least squares with a quadratic constraint," *Numerische Mathematik*, vol. 36, pp. 291–307, 1981.

[51] G. Magoulas, V. Plagianakos, D. Tasoulis, and M. Vrahatis, "Tumor Detection in Colonoscopy Using the Unsupervised k-windows Clustering Algorithm and Neural Networks," in *Fourth European Symposium on Biomedical Engineering*, Patras, Greece, June 2004, pp. 25–27.

[52] A. Hoover, V. Kouznetsova, and M. Goldbaum, "Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response," *IEEE Transactions on Medical Imaging*, vol. 19, pp. 203–210, March 2000.

[53] S. Chaudhuri, S. Chatterjee, N. Katz, M. Nelson, and M. Gold-baum, "Detection of blood vessels in retinal images using two-dimensional matched filters," *IEEE Transactions on Medical Imaging*, vol. 8, no. 3, pp. 263–269, September 1989.

[54] L. Najman and M. Schmitt, "Geodesic Saliency of Watershed Contours and Hierarchical Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 12, pp. 1163–1173, December 1996.

[55] V. Grau, A. Mewes, M. Alcaniz, R. Kikinis, and S. Warfield, "Improved watershed transform for medical image segmentation using prior information," *IEEE Transactions on Medical Imaging*, vol. 23, no. 4, pp. 447–458, April 2004.

[56] H. Nguyen, M. Worring, and R. van den Boomgaard, "Watersnakes: Energy-Driven Watershed Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 3, pp. 330–342, March 2003.

[57] A. Bieniek and A. Moga, "An efficient watershed algorithm based on connected components," *Pattern Recognition*, vol. 33, no. 6, pp. 907–916, June 2000.

[58] L. Vincent and P. Soille, "Watersheds in Digital Spaces: An Efficient Algorithm based on Immersion Simulations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 583–589, June 1991.

[59] L. Vincent, "Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms," *IEEE Transactions on Image Processing*, vol. 2, no. 2, pp. 176–201, April 1993.

[60] A. Fitzgibbon, M. Pilu, and R. Fisher, "Direct least-squares fitting of ellipses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 476–480, May 1999.

[61] G. Klanderman and W. Rucklidge, "Comparing Images Using the Hausdorff Distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[62] C. D.S. and L. Z.K., "A novel approach to detect and correct specular reflectioned face region in color image," in *Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance*, July 2003, pp. 7 – 12.

[63] Y.-P. Tan, D. Saur, S. Kulkarni, and P. Ramadge, "Rapid Estimation of Camera Motion from Compressed Video with Application to Video Annotation," *IEEE Trans. Circuits Syst. Video Technol*, vol. 10, pp. 133–146, Feburary 2000.

[64] J. Kim, H. Chang, J. Kim, and H. Kim, "Efficient camera motion characterization for MPEG video indexing," in *IEEE ICME*, 2000, pp. 1171–1174.

[65] A. Dante and M. Brookes, "Precise Real-Time Outlier Removal from Motion Vector Fields For 3D Reconstruction," in *IEEE ICIP*, Barcelona, 2003, pp. 393–396.

[66] A. Akutsu, Y. Tonomura, H. Hashimoto, and Y. Ohba, "Video indexing using motion vectors," in *SPIE*, 1992, pp. 1522–1530.

[67] B. P., G. M., and G. F., "A unified approach to shot change detection and camera motion characterization," in *IEEE Trans Circuits and Syst. for Video Technol.*, 1999, pp. 1030–1044.

[68] X. Cao and P. N. Suganthan, "Video Shot Motion Characterization based on Hierarchical Overlapped Growing Neural Gas Networks," in *Multimedia Systems*, October 2003, pp. 378–385.

[69] H. Yi, D. Rajan, and L.-T. Chia, "Automatic Extraction of Motion Trajectories in Compressed Sports Videos," in *ACM Multimedia 2004*, New York, NY, October 2004.

BIOGRAPHICAL STATEMENT

Sae Hwang was born on April 24, 1972 in South Korea. In 1995, he received the B.E in Civil Engineering from Chung-Ang University, Seoul, Korea. As soon as he finished his B.E., he served as a first lieutenant in Korean Army for 3 years. After his military service, he received the M.S. in Computer Science at Texas A&M - CC, TX in 2002.

In spring 2003, he entered the Doctoral program in Computer Science and Engineering at the University of Texas at Arlington. Since then, he has been a member of Multimedia Information Group (MIG) and Endoscopic Multimedia Information System (EMIS) working with Dr. JungHwan Oh. During his study, he has worked as a teaching assistant & a research assistant.

His primary research interests include image / video signal processing, computer vision, medical imaging, and multimedia mining. In addition to academic research, his study includes the development of software application achieving a proper balance between academic research and industrial development His work also includes techniques for multimedia data representation and indexing structure, and summary.

Mr. Hwang is a student member of the IEEE and the ACM.