Ensuring High-Quality Colonoscopy by Reducing Polyp Miss-Rates

by

Nima Tajbakhsh

A Dissertation Presented in Partial Fulfillment
of the Requirement for the Degree
Doctor of Philosophy

Approved April 2015 by the
Graduate Supervisory Committee:

Jianming Liang, Chair
Robert Greenes
Matthew Scotch

ARIZONA STATE UNIVERSITY

May 2015

ABSTRACT

Colorectal cancer is the second-highest cause of cancer-related deaths in the United States with approximately 50,000 estimated deaths in 2015. The advanced stages of colorectal cancer has a poor five-year survival rate of 10%, whereas the diagnosis in early stages of development has showed a more favorable five-year survival rate of 90%. Early diagnosis of colorectal cancer is achievable if colorectal polyps, a possible precursor to cancer, are detected and removed before developing into malignancy.

The preferred method for polyp detection and removal is optical colonoscopy. A colonoscopic procedure consists of two phases: (1) insertion phase during which a flexible endoscope (a flexible tube with a tiny video camera at the tip) is advanced via the anus and then gradually to the end of the colon–called the cecum, and (2) withdrawal phase during which the endoscope is gradually withdrawn while colonoscopists examine the colon wall to find and remove polyps. Colonoscopy is an effective procedure and has led to a significant decline in the incidence and mortality of colon cancer. However, despite many screening and therapeutic advantages, 1 out of every 4 polyps and 1 out of 13 colon cancers are missed during colonoscopy.

There are many factors that contribute to missed polyps and cancers including poor colon preparation, inadequate navigational skills, and fatigue. Poor colon preparation results in a substantial portion of colon covered with fecal content, hindering a careful examination of the colon. Inadequate navigational skills can prevent a colonoscopist from examining hard-to-reach regions of the colon that may contain a polyp. Fatigue can manifest itself in the performance of a colonoscopist by decreasing diligence and vigilance during procedures. Lack of vigilance may prevent a colonoscopist from detecting the polyps that briefly appear in the colonoscopy videos. Lack of diligence may result in hasty examination of the colon that is likely to miss polyps and lesions.

To reduce polyp and cancer miss rates, this research presents a quality assurance system with 3 components. The first component is an automatic polyp detection system that highlights the regions with suspected polyps in colonoscopy videos. The goal is to encourage more vigilance during procedures. The suggested polyp detection system consists of several novel modules: (1) a new patch descriptor that characterizes image appearance around boundaries more accurately and more efficiently than widely-used patch descriptors such HoG, LBP, and Daisy; (2) A 2-stage classification framework that is able to enhance low level image features prior to classification. Unlike the traditional way of image classification where a single patch undergoes the processing pipeline, our system fuses the information extracted from a pair of patches for more accurate edge classification; (3) a new vote accumulation scheme that robustly localizes objects with curvy boundaries in fragmented edge maps. Our voting scheme produces a probabilistic output for each polyp candidate but unlike the existing methods (e.g., Hough transform) does not require any predefined parametric model of the object of interest; (4) and a unique three-way image representation coupled with convolutional neural networks (CNNs) for classifying the polyp candidates. Our image representation efficiently captures a variety of features such as color, texture, shape, and temporal information and significantly improves the performance of the subsequent CNNs for candidate classification. This contrasts with the exiting methods that mainly rely on a subset of the above image features for polyp detection. Furthermore, this research is the first to investigate the use of CNNs for polyp detection in colonoscopy videos.

The second component of our quality assurance system is an automatic image quality assessment for colonoscopy. The goal is to encourage more diligence during procedures by warning against hasty and low quality colon examination. We detect a low quality colon examination by identifying a number of consecutive non-informative

frames in videos. We base our methodology for detecting non-informative frames on two key observations: (1) non-informative frames most often show an unrecognizable scene with few details and blurry edges and thus their information can be locally compressed in a few Discrete Cosine Transform (DCT) coefficients; however, informative images include much more details and their information content cannot be summarized by a small subset of DCT coefficients; (2) information content is spread all over the image in the case of informative frames, whereas in non-informative frames, depending on image artifacts and degradation factors, details may appear in only a few regions. We use the former observation in designing our global features and the latter in designing our local image features. We demonstrated that the suggested new features are superior to the existing features based on wavelet and Fourier transforms.

The third component of our quality assurance system is a 3D visualization system. The goal is to provide colonoscopists with feedback about the regions of the colon that have remained unexamined during colonoscopy, thereby helping them improve their navigational skills. The suggested system is based on a new 3D reconstruction algorithm that combines depth and position information for 3D reconstruction. We propose to use a depth camera and a tracking sensor to obtain depth and position information. Our system contrasts with the existing works where the depth and position information are unreliably estimated from the colonoscopy frames. We conducted a use case experiment, demonstrating that the suggested 3D visualization system can determine the unseen regions of the navigated environment. However, due to technology limitations, we were not able to evaluate our 3D visualization system using a phantom model of the colon.

To my departed father,

who would have loved to see this moment

ACKNOWLEDGEMENTS

First and foremost, I would like to extend my deepest gratitude to my advisor, Dr. Jianming Liang, for his guidance and support during my PhD journey. His visions were a source of inspiration and encouragement in the lowest periods of my PhD life. I am very grateful to him for his scientific advice, knowledge, and many insightful discussions and suggestions. I would also like to extend a very special thanks to Dr. Suryakanth Gurudu for his continuous effort in providing us with the video data. Without his support, I would never have been able to complete this project. I would further like to thank the members of my PhD committee, Prof. Greenes and Dr. Scotch, for their helpful career advice and suggestions in general.

I am very grateful to the former members of the imaging informatics laboratory, Hong Wu and Wenzhe Xue, who patiently helped me with some of the projects I did in the first two years of my PhD. I am also very grateful to several undergraduate students at ASU, Sarah Fallah-Adl, Saiswathi Javangula, Ireen Khan, Kamran Bodushev, and Tracy Phan, who tremendously helped me with creating ground truth for the collected videos. Their contributions were truly significant.

I would like to extend my special gratitude to Nikou Hesari, who patiently supported me in the hardest periods of my PhD journey. She was a constant source of encouragement and a great friend, for which I will be forever indebted to her. I wish her a productive career and a happy life ahead.

Last, but not the least, I want to thank my loving family—my mother, Dr. Akhtar Khosravi; my sister, Dr. Neda Tajbakhsh; and my late father, Dr. Iraj Tajbakhsh, for their immense love, support and care. Without them, I would never have been able to complete this or any other endeavor in life.

TABLE OF CONTENTS

Page

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER

1 BACKGROUND 1

1.1 Colorectal Cancer 1

1.2 Colonoscopy 1

1.3 Polyp Miss-Rates 2

1.4 Cancer Miss-Rates 3

1.5 Why Are Polyps and Cancers Missed? 3

1.6 Quality Guidelines 4

1.7 Effectiveness of Quality Metrics 5

1.8 The Technology Gap 6

1.9 Proposed Solutions 7

1.10 Organization of the Dissertation 9

2 AUTOMATIC POLYP DETECTION 10

2.1 Related Work 10

2.2 Proposed Polyp Detection System 12

2.2.1 Stage 1: Candidate Generation 14

2.2.2 Stage 2: Candidate Classification 30

2.3 Data 36

2.3.1 CVC-ColonDB 36

2.3.2 ASU-Mayo Clinic Polyp Database 37

2.4 Experiments 38

CHAPTER Page

2.4.1 Model Tuning 38
2.4.2 Model Evaluation 46
2.4.3 Performance Comparison 49
2.5 Conclusion 53
3 AUTOMATIC VIDEO QUALITY ASSESSMENT 54
3.1 Related Work 54
3.2 Proposed Video Quality Assessment Method 55
3.3 Data 58
3.4 Performance Evaluation and Comparison 59
3.5 Visual Feedback 60
3.6 Conclusion 61
4 3D VISUALIZATION of the COLON 64
4.1 Related Work 64
4.2 Proposed 3D Visualization System 66
4.2.1 Tracking Sensor 66
4.2.2 Depth Camera 67
4.2.3 3D Reconstruction Pipeline 68
4.2.4 A use case 75
4.2.5 Extension to Colonoscopy 76
4.3 Conclusion 78
5 DISCUSSION 80
5.1 Contributions 80
5.2 Limitations and Future Work 82

CHAPTER Page

REFERENCES 86

APPENDIX

A CANDIDATE GENERATION METHOD 92

B EDGE CLASSIFICATION SCHEME 94

C 3D RECONSTRUCTION PIPELINE 97

LIST OF TABLES

Table Page

2.1 Recent Polyp Detection Methods Designed for Colonoscopy Videos 12

2.2 R^2 Coefficient for the Regression Models Constructed under Different Configurations of the Voting Scheme. 44

2.3 Performance Comparison Between Our System and the Recent Polyp Detection Methods Designed for Optical Colonoscopy. 51

2.4 Performance Comparison Based on *CVC-ColonDB*. 52

4.1 Parameters of the Senz3D Camera Retrieved from the API. 71

LIST OF FIGURES

Figure Page

1.1 Causes of Polyp Miss-rates and the Solutions Explored in This Dissertations. . . . 9

2.1 Significant Color Variation among Different Instances of the Same Polyp Caused by Varying Lighting Conditions. . . . 11

2.2 Non-polyp Structures with Similar Geometric and Shape Features to Those of Polyps. . . . 11

2.3 An Overview of Our Polyp Detection System. . . . 13

2.4 Visual Characteristics of Polyps 15

2.5 Ball Tensor Voting Determines Gradient Orientations More Reliably than the Traditional Method Based on Horizontal and Vertical Image Gradients. . . . 16

2.6 Basis Functions Corresponding to Low Frequency Dct Coefficients. . . . 20

2.7 The Test Stage of the Suggested Classification Scheme. . . . 21

2.8 Pair of Patches at Each Edge Location. . . . 22

2.9 The Training Stage of the Suggested Classification Scheme. . . . 24

2.10 The Generated Voting Map for an Edge Pixel Lying at 135 Degree. . . . 28

2.11 Voting Process. . . . 29

2.12 Using Voting Directions and Map Multiplication Is Essential for Correct Polyp Localization. . . . 31

2.13 Determining a Narrow Band Around a Polyp Candidate. . . . 32

2.14 The Suggested Score Fusion Framework Based on Convolutional Neural Networks. . . . 36

2.15 Examples of Polyps and the Corresponding Ground Truth Images from the Asu-Mayo Clinic Database. . . . 38

Figure Page

2.16 Evaluation of Our Patch Descriptor. 41
2.17 Network Layout for the Deep Convolution Networks. 46
2.18 FROC Analysis of Our Polyp Detection System. 47
2.19 Detection Latency Analysis of Our Polyp Detection System. 49
2.20 Variable Importance Maps Generated by Random Forest Classifiers. . . . 52
3.1 Comparison Between Informative and Non-informative Frames. 56
3.2 Overview of the Suggested Method for Image Informativeness Assessment. 58
3.3 Quantitative and Qualitative Performance Comparison. 62
3.4 Examples of Visual Feedback Generated by Our Quality Assessment System. 63
4.1 The Tracking System We Used in This Study. 67
4.2 A Depth Image Acquired by Creative Senz3D^{TM}. 68
4.3 Moving the Depth Image from a Reference Location to the Target Location by Rotation R Followed by Translation T. 69
4.4 Projecting a Depth Image to a Cloud of Points in 3D. 72
4.5 Cloud Matching Based on the Camera Information. 75
4.6 Different Views of the 3D Model Constructed for a Box. 79

Chapter 1

BACKGROUND

1.1 Colorectal Cancer

Colorectal cancer is the third most commonly diagnosed cancer and the second leading cause of cancer death in the US. The American cancer society estimates that in 2015 approximately 140,000 people will be diagnosed with colorectal cancer, and 50,000 people will die from it in the US Siegel *et al.* (2015). Colorectal cancer most often arises from polyps—abnormal growths inside the colon. However, polyps grow slowly and may take years to turn into cancer. Therefore, early detection of polyps can decrease the incidence and mortality of colon cancer. In fact, while the advanced stages of colorectal cancer has a poor five-year survival rate of 10%, the early diagnosis has showed a more favorable five-year survival rate of 90% Rabeneck *et al.* (2003). Early diagnosis of colorectal cancer is achievable if colorectal polyps are detected and removed before developing into malignancy.

1.2 Colonoscopy

The gold standard screening method for polyp detection and removal is optical colonoscopy Rex *et al.* (1997). A colonoscopic procedure consists of two phases: insertion phase and withdrawal phase. During the insertion phase, a flexible endoscope is advanced via the anus into the rectum and then all the way to the end of the colon (the cecum). Reaching the cecum is of critical importance since a substantial number of polyps reside around the cecum. The endoscope is a flexible tube with a tiny video camera at the tip, which transmits video signals to a large LCD in the operating

room. In the withdrawal phase, colonoscopists gradually withdraw the endoscope while watching the video on the screen. The purpose of the withdrawal phase is to carefully inspect the colon to detect and remove polyps.

1.3 Polyp Miss-Rates

Colonoscopy has been a successful preventive procedure and has led to a 30% decline in the incidence of colon cancer. However, its effectiveness for detecting polyps and cancer is dependent on the quality of the procedure Lieberman *et al.* (2007). A colonoscopy procedure with inadequate quality is likely to miss polyps and cancers. In fact, as evidenced by several clinical studies Rex *et al.* (1997); van Rijn *et al.* (2006); Heresbach *et al.* (2008); Gelder *et al.* (2004), a significant fraction of polyps are missed during colonoscopy. The pioneer study on polyp miss-rate Rex *et al.* (1997) was conducted in the mid 1990s, reporting 0% to 6% miss-rate of adenomas [1] $\geq$1 cm, 12% to 13% for adenomas 6 to 9 mm in size, and 15% to 27% for adenomas $\leq$5 mm in size. Following this study, more clinical trials were launched to measure the fraction of polyps being missed during colonoscopy, among which van Rijn *et al.* (2006) reported 22% pooled miss rate for polyps of all sizes and Heresbach *et al.* (2008) reported 11% miss-rate for adenomas $\geq$ 5 mm. Despite the differences between these clinical trials, they were all similar in that they used colonoscopy as its own gold standard, meaning that each colonoscopy was repeated with a new colonoscopist to identify the missed polyps. Obviously, using colonoscopy as its own gold standard underestimates polyp miss-rates. In response to this drawback, Gelder *et al.* (2004) measured the adenoma miss rate of colonoscopy, using virtual colonography as the reference standard, and found out that 17% of adenomas $\geq$ 1 cm is missed during the procedures. All these

[1] Adenomas or pre-cancerous polyps are a particular kind of polyps that will develop into malignancy and cause colorectal cancer. Note that, not all polyps lead to cancer.

findings on adenoma miss-rates may lead to the following question *"Is colonoscopy for colorectal cancer prevention fulfilling the promise?"* Robertson (2010).

1.4 Cancer Miss-Rates

The concerning issue of missed colorectal cancer has recently been studied by researchers at the University of Toronto Bressler *et al.* (2007), where they reviewed data from 12,487 colorectal cancer patients who had undergone colonoscopies within 3 years before their diagnoses. Patients with new or missed cancers were those whose most recent colonoscopy was 6 to 36 months before diagnosis—their last colonoscopy was negative with no suspicious findings. Their study revealed a cancer miss-rate of 4% to 6%, indicating that between 4% to 6% of the colonoscopies that could potentially detect colorectal cancer had failed to do so. Although biological factors that lead to rapid polyp progression is an additional explanation for the missed cancers, the existing evidences on polyp miss-rate are compelling enough to identify the missed polyps as the main culprit for the occurrence of interval cancer. Such speculation is also supported by another study Robertson *et al.* (2008) on 9167 patients with adenomas, where missed or incomplete resections of polyps was determined to be the cause of 58 missed cancers.

1.5 Why Are Polyps and Cancers Missed?

Low-quality colonoscopy can lead to miss detection of polyps and cancers. There are many factors the can degrade the quality of colonoscopy including poor colon preparation Lebwohl *et al.* (2011), inadequate navigational skills Rex (2000), and fatigue Lee *et al.* (2011). In the following, we focus on fatigue and inadequate navigational skills, which are the main scope of this dissertation.

Fatigue. Recent studies have identified fatigue as a major contributing factor to missed polyps and cancers. Typically, colonoscopists are assigned an entire day at a time to work in the endoscopy laboratory. However, performance of a colonoscopist is not the same over the entire day. For example, a recent study Lee *et al.* (2011) has shown that morning colonoscopy detects significantly more polyps than the procedures performed later in the day, and more specifically, each elapsed hour in the day may be associated with a 4.6% reduction in polyp detection rates. Fatigue manifests itself in the performance of colonoscopists by decreasing their diligence and vigilance during the procedures. Lack of diligence results in hasty colon examination and lack of vigilance can lead to superficial inspection of colonoscopy videos, which both can lead to the miss detection of polyps

Navigational skills. The key to a complete colon examination is to master navigational skills Hewett *et al.* (2010). Colonoscopists vary in their rate of their skill acquisition Mohamed *et al.* (2011) and it often take years for colonoscopists to procure the navigational skills Raman and Donnon (2008) such as maintenance of luminal view, torque steering, lumen distention, and using position change. Inadequate navigational skills can lead to an incomplete colonoscopy and that results in the miss detection of polyps that reside in the unseen regions of the colon.

1.6 Quality Guidelines

In response to the increasing concerns for missed polyps and cancers, American Society for Gastrointestinal Endoscopy (ASGE) established a number of guidelines for a quality colonoscopy in 2002 Rex *et al.* (2002) and a set of adjusted guidelines in 2006 Rex (2006). These measures that aim to improve the quality of colonoscopy can be categorized as pre-procedure quality indicators, intra-procedure quality indicators, and post-procedure quality indicators. The pre-procedure quality measures

are mainly to ensure that there exists a valid indication for colonoscopy, and also necessitate a well-prepared colon before performing the procedure. The intra-procedure quality measures mainly recommend colonoscopists to intubate the cecum and also to maintain a minimum of 6-minute withdrawal phase. And the post-procedure quality measures are mostly concerned with perforation and other complications of colonoscopy. According to the quality guidelines, colonoscopy reports should include withdrawal time, photographic evidence of cecal intubation, and score of bowel preparation.

1.7 Effectiveness of Quality Metrics

It has been shown that strict adherence to these quality guidelines leads to a higher quality of colonoscopy with improved adenoma detection rate. For example, a recent study Lee *et al.* (2012) reports high-quality colonoscopies by ensuring that a high cecal intubation rate, adenoma detection rate, and low adverse event rates. Another clinical trial Saini *et al.* (2009) conducted by eMerge Health Solutions in 2007 and 2008 revealed improvement in adenoma detection from 26.6% to 31.8% in men and from 19.8% to 28.7% in women after real-time visual feedback for colonoscopy withdrawal times was provided to the participating group of colonoscopists. Another study by researchers at the University of Michigan Health System Frandzel (2007) estimates a 20% increase in detection of large polyps ($\geq$1 cm) after colonoscopists received a $50 bonus per colonoscopy for maintaining an average withdrawal time of 8 minutes or more. These findings suggest that faithful adherence to quality measures can potentially improve colonoscopy outcomes.

Some European countries have also implemented quality assurance for colonoscopy. In Germany, quality indicators along with findings and complications need to be documented in colonoscopy reports, otherwise there is no reimbursement ("no cecum

intubation, no money") Birkner *et al.* (2005). In France, quality assurance of endoscopic procedures has been a legal obligation for healthcare professionals since 2004. The UK has an advanced quality assurance program for screening colonoscopy. The status of accredited screening colonoscopists is reviewed on annual basis. To hold the accreditation, endoscopists must perform more than 150 screening colonoscopies a year with more than a 90% cecal intubation rate, bowel preparation should be adequate in more than 90% of patients, and adenoma detection rates and withdrawal time must meet the standard. Despite structural implementation and monitoring of quality indicators in European countries, such quality guidelines merely serve as recommendations in the U.S. with no official process to oversee the quality of performed colonoscopies.

1.8 The Technology Gap

Although quality guidelines have proven to improve polyp detection rates, they are not sufficient to ensure a high quality procedure. For example, one intra-procedure quality guideline recommends colonoscopists to maintain a minimum withdrawal time of 6 minutes. While this metric encourages more diligence during procedures, it cannot guarantee a quality and thorough colon inspection. This is because a colonoscopist may spend a large amount of time in one segment of the colon (for removing a polyp or getting a biopsy) but performs a quick examination in the other parts. In doing so, the colonoscopists meet the minimum examination time, but they have not achieved the expected quality. Another important intra-procedure quality guideline recommends colonoscopists to videotape or photo-document the end of the colon, which is where a substantial number of polyps reside. However, this quality guideline only ensures the examination of the end segment of the colon and not a careful examination of the entire colon. Clearly, the existing intra-procedure quality guidelines

are limited in value and are unable to ensure a high-quality colonoscopy. Therefore, there is a technology gap, demanding innovative technologies to ensure the quality of colonoscopy.

1.9 Proposed Solutions

We aim at ensuring high-quality colonoscopy by reducing polyp miss-rates. For this purpose, we categorize the missed polyps into three categories and propose a solution for each category of missed polyps. Figure 1.1 illustrates our categorization of the polyps missed during colonoscopy and the suggested solutions.

The first category (**C1**) contains polyps that appear in the videos clearly but get overlooked by colonoscopists due to lack of vigilance. Examples include the small polyps that appear in similar color and texture with the surrounding tissue or the polyps that appear briefly in the borders of the colonoscopic view. The second category (**C2**) contains polyps that are located in the regions of the colon that have been visited by the colonoscopy camera but with inadequate diligence. Colonoscopy videos captured during a hasty examination result in a large number of blurred and non-informative frames in which polyps either partially or indistinctly appear and thus unrecognizable by colonoscopists. The third group (**C3**) contains the polyps that are located in regions of the colon that have not been visited by the colonoscopy camera due to inadequate navigational skills. Examples include polyps that tend to remain hidden behind the folds or polyps that reside in regions that are hard-to-reach by colonoscopists.

We propose a technology for each category of missed polyps:

1. **Automatic polyp detection for category C1:** To detect as many polyps as possible, colonoscopists must vigilantly watch the colonoscopy video on the screen. However, fatigue may lead to a deviation from an attentive examination.

When a polyp appears in the colonoscopy video, the colonoscopist, if inattentive, may overlook the polyp, even if it is clearly recognizable in the video. Our technology aims to encourage vigilance by automatically detecting the polyps in the live video and highlighting their locations on the screen for colonoscopists. Continual highlights during a procedure will engage colonoscopists attention and assist them with the task of polyp detection.

2. **Automatic video quality assessment for category C2:** To achieve high-quality colonoscopy, colonoscopist must diligently navigate the scope in the colon. However, when exhausted, the colonoscopist tends to rush the procedure by withdrawing the scope quickly in an effort to finish the procedure. A rapid withdrawal of the scope yields a video with many blurry and out-of-focus images. If any polyps appear in such poor-quality frames, they might not be easily recognizable, resulting in missed detections. Therefore, our technology aims to automatically detect lack of diligence during the procedure by identifying a series of uninformative frames (i.e., blurred and unfocused images), thereby warning colonoscopists against a hasty withdrawal.

3. **3D colon visualization for category C3:** To search the entire colon for polyps, colonoscopists must master navigational skills such as maintenance of luminal view and torque steering. However, these skills are accumulated over years of practice. With sub-optimal navigational skills, some regions in the colon are likely to be missed during the colonoscopy. Our technology aims to identify parts of the colon that remain unseen by building a 3D model of the colon from the colonoscopy video. Such feedback should be provided in a real-time fashion, allowing colonoscopists to go back and re-examine the missed parts of the colon. This research will explore the feasibility of a system for 3D visualization of the

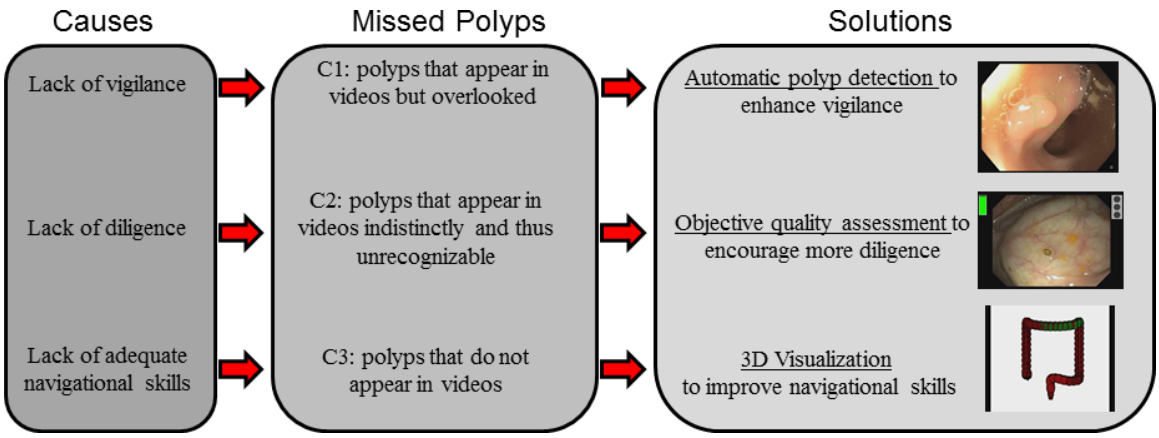


Figure 1.1: Causes of polyp miss-rates and the solutions explored in this dissertations.

colon.

1.10 Organization of the Dissertation

The rest of the dissertation is organized as follows: Chapter 2 describes our polyp detection system for colonoscopy videos. Chapter 3 describes our suggested video quality assessment system. Chapter 4 presents a use case for a 3D visualization system based on a depth camera and a tracking sensor. Chapter 5 concludes this dissertation by outlining the contributions and discussing the limitations and future directions.

Chapter 2

AUTOMATIC POLYP DETECTION

Computer-aided polyp detection may assist colonoscopists with the task of polyp detection. The goal is to provide feedback about the locations of polyps in the colon, helping colonoscopists stay alert and vigilant during colonoscopy.

This chapter proceeds with a summary of the existing methods for polyp detection in colonoscopy videos followed by a discussion of their limitations and strengths. It then presents our suggested polyp detection system. Next, our polyp database collected at Mayo Clinic in Arizona is described. Finally, experimental results are presented.

2.1 Related Work

The early work of Karkanis *et al.* (2003) was published in 2003 where color wavelet features coupled with a sliding window was used for detecting polyps in colonoscopy images. This article inspired more research Iakovidis *et al.* (2005); Alexandre *et al.* (2008); Ameling *et al.* (2009); Cheng *et al.* (2011) on polyp detection using texture and color descriptors. However, these methods are limited by partial texture visibility and large color variations among polyps. The former is caused by the relatively large distance between the polyps and the single-focus camera and the latter is caused by varying lighting conditions. The reliability of these methods was also questioned in Bernal *et al.* (2012). Figure 2.1 shows variation in color and appearance of the same polyp in a sequence of frames.

The other category of techniques for polyp detection employed shape, spatio-temporal, and appearance features. Hwang *et al.* (2007) suggested elliptical shape features for detecting the shots of polyps in colonoscopy videos. The suggested ap-

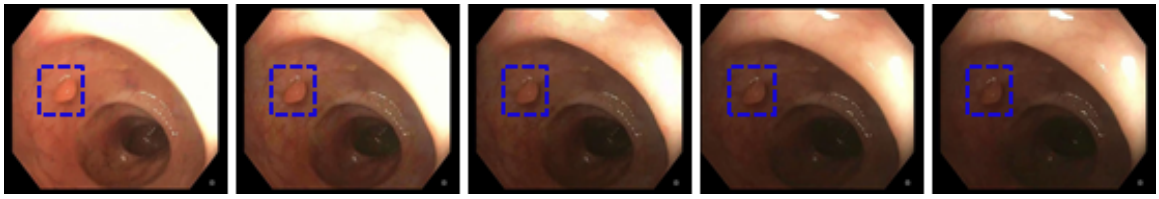


Figure 2.1: Significant color variation among different instances of the same polyp caused by varying lighting conditions.

Figure 2.2: Non-polyp structures with similar geometric and shape features to those of polyps.

proach, however, did not consider image context when extracting the shape clues, allowing for false positive detections around other elliptical structures in the complex endoluminal scenes. Figure 2.2 shows examples of non-polyp structures with similar geometric features to that of polyps. Bernal *et al.* (2012) employed valley information and a region growing approach to find polyps in colonoscopy images. However, as acknowledged by the authors, this method may result in false detections particularly around vascular structures. Finally, the spatio-temporal features suggested by Park *et al.* (2012) is suitable only for the off-line processing of colonoscopy videos since it requires information from the past and future frames for polyp localization at the current frame. A summary of the recent polyp detection methods for colonoscopy videos is presented in Table 2.1.

Polyp detection and classification has also been considered in CT colongraphy Zhao *et al.* (2006); Lee *et al.* (2011); Ong and Seghouane (2011); Zhu *et al.* (2011); Suzuki *et al.* (2010); Van Ravesteijn *et al.* (2010); van Wijk *et al.* (2010), wireless

Table 2.1: Recent polyp detection methods designed for colonoscopy videos (SVM: support vector machines, MI: mutual information).

Author	**Feature**	**Classifier**
Wang *et al.* (2013)	Edge cross section profile	SVM
Bernal *et al.* (2012)	Valley information	—
Park *et al.* (2012)	Temporal and appearance features	Conditional random field
Cheng *et al.* (2011)	Texture features	SVM
Ameling *et al.* (2009)	Texture and color features	SVM
Hwang *et al.* (2007)	Temporal and elliptical shape features	distance based on MI

capsule endoscopy Silva *et al.* (2013); Figueiredo *et al.* (2013); Hwang and Celebi (2010), and narrow band imaging Kwitt *et al.* (2011); Gross *et al.* (2012); Tamaki *et al.* (2013). However, the challenges posed by these imaging modalities differ from that of colonoscopy. To design a polyp detection system for colonoscopy, one needs to consider the effects of varying lighting conditions, specular reflections, spontaneous colon deformation, and varying view angles of the camera. However, such challenges do not or partially apply to the other imaging modalities. For instance, while texture is not reliable for polyp detection in colonoscopy, the pit patterns of polyps are heavily used in narrow-band imaging; or while shape and curvature clues have been successfully used for polyp detection in CT colonography, they are misleading in the complex colonoscopy images if not combined with the context clues.

2.2 Proposed Polyp Detection System

Our computer-aided polyp detection method has two stages. The first stage utilizes geometric features to reliably generate polyp candidates and the second stage employs a comprehensive set of deep features to effectively remove false candidates. Our system is motivated by the fact that while shape features are reliably available and are descriptive of curvilinear heads of polyps, they lack sufficient discriminative

power that comes with color, texture and temporal features. In Figure 2.3, we show a schematic overview of the suggested method.

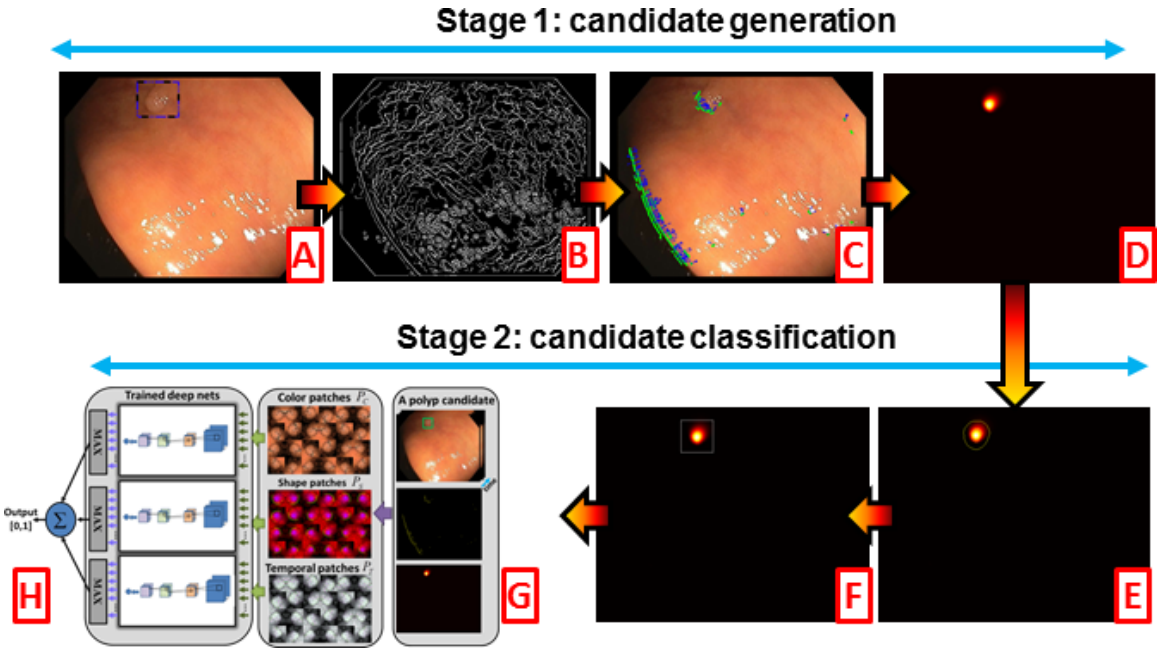


Figure 2.3: An overview of our polyp detection system. Our system consists of 2 stages: candidate generation and classification. Given a colonoscopy frame (A), we obtain a crude set of edge pixels (B) (see Section 2.2.1.1). We then refine this edge map using a feature extraction (see Section 2.2.1.2) and classification scheme (see Section 2.2.1.3) where the goal is to remove as many non-polyp boundary pixels as possible (C). The geometric features of the retained edges are then captured through a voting scheme (see Section 2.2.1.4), generating a voting map whose maximum indicates the location of a polyp candidate (D). In the second stage, a band and then a bounding box (see Section 2.2.2.1) is estimated for each generated candidate (E,F). Given the candidate and the corresponding bounding box, a set of convolution neural networks (G)—each specialized in one type of features—are applied in the vicinity of the candidate (see Section 2.2.2.2). Finally, the outcomes of CNNs are aggregated to generate a confidence value (H) for the given polyp candidate.

2.2.1 Stage 1: Candidate Generation

Our candidate generation method exploits the following two properties:

1. Polyps have distinct appearance across their boundaries. In Figure 2.4(a), we illustrate this property, where hundred thousands oriented image patches around polyps, vessels, lumen, and specular reflections are averaged and compared. Figure 2.2 shows how these patches are extracted from one of the used images. As seen in Figure 2.1, average polyp appearance (top-left) is distinct. Factors that contribute to the distinct appearance around polyp boundaries include the depth contrast between the polyp side and background side of the boundaries, color contrast between polyps and the colon surface due to biological changes during the development of polyps, lighting conditions, and polyps' levels of protrusion. We use this property of polyps in designing our novel image characterization (Section 2.2.1.2) and edge classification schemes (Section 2.2.1.3).

2. Polyps, irrespective of their morphology and varying levels of protrusion, feature at least one curvilinear head at their boundaries. In Figure 2.4(c), we illustrate this property by highlighting the curvilinear heads of several polyps. We use this property of polyps in designing our voting scheme (Section 2.2.1.4) that localizes polyps by detecting objects with curvy boundaries.

As shown in Figure 2.3, our candidate generation method consists of three steps: (1) constructing an edge map for an input image, (2) refining the edge map by classifying every edge pixel into polyp and non-polyp categories using boundary appearance (the first property of polyps), (3) localizing polyp candidates from the refined edge maps using shape information (the second property of polyps). In the following, we describe each stage of the suggested candidate generation method.

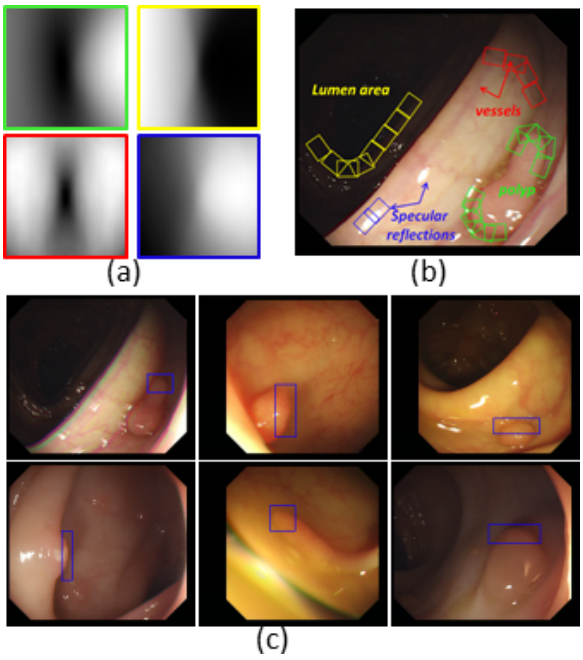


Figure 2.4: Visual characteristics of polyps. (a) From left to right: average appearance of polyps, lumen, vessels, and specular reflection across thousands of image patches. Polyp boundary has a distinct appearance. (b) Illustrating how the patches have been collected from images. (c) Despite varying morphology, polyps feature a curvy segment in their boundaries.

2.2.1.1 Constructing Edge Maps

We apply Canny's method on the three color channels of the input images to extract as many edges as possible. Once an edge map is constructed, we estimate gradient orientations for all the edge pixels in the map. We later use the gradient orientations to extract oriented patches around the edge pixels. Canny's algorithm computes gradient directions based on the local image gradients in horizontal and vertical directions; however, such estimations are often not accurate, leading to a non-smooth gradient direction map. Alternatively, we estimate gradient orientations by performing ball tensor voting Mordohai and Medioni (2007). A complete description

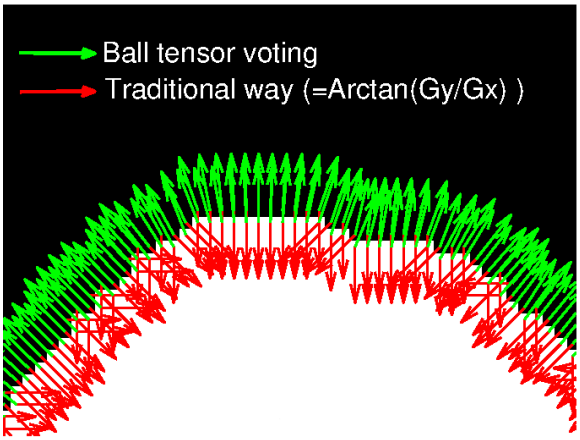


Figure 2.5: Ball tensor voting determines gradient orientations more reliably than the traditional method based on horizontal and vertical image gradients. For better visualization, we have drawn the gradient directions so that they point outward for ball tensor voting and inward for the traditional way.

of tensor voting is beyond the scope of this dissertation, but we intuitively explain why tensor voting yields more accurate estimations of gradient orientations than the traditional method based on image gradients.

In ball tensor voting, the gradient orientation of an edge pixel is determined using the arrangements of the surrounding edge pixels such that the continuation of gradient orientation is maintained. In fact, the locations of nearby edges determine the gradient orientation of the central pixel. It is therefore unlikely to obtain an inconsistent or non-smooth orientation map. This is in contrast to the traditional methods (e.g., Roberts, Prewitt, Sobel) that discard the configuration of the surrounding edge pixels. Figure 2.5 compares the gradient orientations obtained using ball voting and the traditional way based on horizontal and vertical gradients. As seen, the gradient orientations are more reliably estimated using ball tensor voting.

2.2.1.2 Feature Extraction.

The goal of feature extraction is to capture intensity variation patterns in an image patch around each edge pixel. The desired feature extraction method must meet three major requirements: (1) it must be fast in order to handle a large volume of input patches from the edge detection stage; (2) it must provide high level of illumination invariance, since in a colonoscopy, the source of light moves along with the camera, causing the same segment of a boundary to appear with varying contrast in a number of consecutive frames; (3) it must provide rotation invariance against the edge orientations because the essential information do not lie along the edge directions but across the edges.

Our patch descriptor begins with extracting an oriented patch around an edge pixel such that the edge segment appears vertically in the middle of the patch. This presentation allows us to characterize intensity variation patterns across the edges independent of their orientations. We then form sub-patches of size $n \times m$ all over an extracted patch with 50% overlap along horizontal and vertical directions. Each sub-patch is then averaged vertically, resulting in a 1D intensity signal S, embedding intensity variations along the horizontal axis. To obtain a compact and robust presentation of intensity variations, we then apply a 1D discrete cosine transform (DCT) to the extracted signal:

$$C_k = \frac{2}{n} w(k) \sum_{i=0}^{n-1} S[i] \cos(\frac{2i+1}{2n}\pi k) \tag{2.1}$$

where

$$w(k) = 1/\sqrt{2}, k = 0 \text{ and } w(k) = 1, 1 \leq k \leq n-1.$$

DCT has a strong energy compaction property, allowing the entire spatial infor-

mation to be summarized in a few coefficients. However, such a compact presentation of the signal is not robust against illumination changes. A constant change in the intensity of the pixels in a patch directly appears in the DC coefficient, and illumination scaling affects both the DC and AC coefficients. To achieve invariance against constant illumination changes, we discard the DC component. To achieve invariance against linear intensity changes, we normalize the AC components using their L^2-norm. However, this normalization scheme is not efficient given that we are interested in only a few of the AC components. We therefore compute the first few AC coefficients from each sub-patch and use their L^2-norm for normalization, achieving a significant performance speedup. Finally, the coefficients selected from each sub-patch are concatenated to form a feature vector for the extracted patch.

The suggested patch descriptor has four advantages. First, it is fast because compressing each sub-patch into a 1D signal eliminates the need for expensive 2D DCT, and that only a few DCT coefficients are computed from each intensity signal. Second, due to the normalization treatment applied to the DCT coefficients, our descriptor achieves invariance to linear illumination changes and partial tolerance against nonlinear illumination variations over the entire patch, particularly if the nonlinear change can be decomposed to a set of linear illumination changes on the local sub-patches. Third, our descriptor provides a rotation invariant presentation of the intensity variation patterns due to the consistent appearance of the edge segments in the middle of oriented patches. Fourth, our descriptor handles small positional changes, which is essential to cope with patch variations due to edge mislocalization. In practice, spurious edges around polyp boundaries and Gaussian smoothing prior to Canny edge detector can cause inaccurate edge localization where an edge pixel is found a few pixels away from the actual location of the boundary. It is important for a patch descriptor to provide a consistent image presentation in the presence of

such positional changes. We decrease positional variability by selecting and averaging overlapping sub-patches in both horizontal and vertical directions.

DCT is more suitable for the suggested patch descriptor than the other transforms such as Discrete Sine Transform (DST) and Discrete Fourier Transform (DFT). This is because DCT enables energy compaction for a wider range of intensity signals than the other two transforms. More specifically, DFT assumes that the intensity signal S is a part of a periodic signal; therefore, if the intensity values at both ends of the signal are not equal ($S[0] \neq S[n-1]$), S will appear as a part of non-continuous periodic function to DFT. As such, one can expect large high-frequency Fourier coefficients, which prevents the information (energy) of the signal to be compressed in a few Fourier coefficients. Large high frequency components can also appear in the case of DST, if the intensity signals have non-zero values at their both ends ($S[0] \neq 0, S[n-1] \neq 0$). This constraint requires the corresponding sub-patches to appear dark in the left and right boarders, a constraint that is often not met in practice. In contrast, DCT relaxes these constraints, requiring only smooth behavior at both ends of the intensity signals. This is a more practical assumption because the intensity signals are computed by averaging the corresponding sub-patches, which smoothes out the noise and abrupt changes in intensity values.

We also prefer DCT over Discrete Wavelet Transform (DWT) because it allows for intuitive feature selection and less computational burden. DCT requires only one parameter: the number of selected coefficients from each sub-patch, which can be determined both experimentally and intuitively. In Figure 2.6, we show the first 4 DCT basis functions that are used for feature computation. As seen, the first basis function corresponds to the DC component, the second one measures whether the intensity signal S is monotonically decreasing (increasing) or not, the third one measure the similarity of the intensity signal against a valley (ridge), and finally the fourth one

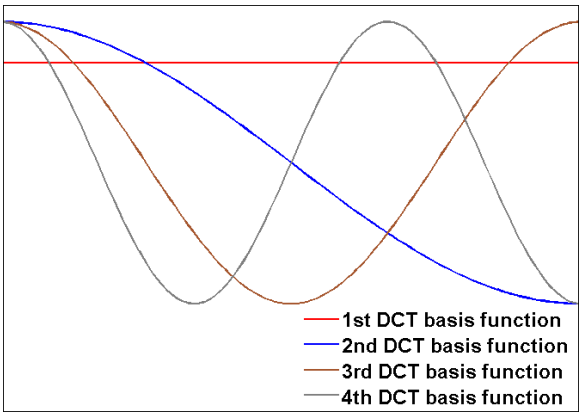


Figure 2.6: Basis functions corresponding to low frequency DCT coefficients.

checks for the existence of both a valley and a ridge in the signal. The higher order basis functions are suitable for detecting higher frequency variations in the intensity signal; however, such components may not be reliable as they are susceptible to noise and degradation factors in the images. Therefore, one can intuitively set the number of desired coefficients without resorting to more complicated feature selection algorithms. In contrast, DWT requires the tuning of more parameters such as the choice of wavelet function or the number of decomposition levels, resulting in a large number of wavelet coefficients, which demands an appropriate feature selection mechanism to produce an efficient image presentation.

2.2.1.3 Edge Classification.

The purpose of edge classification is two-fold: (1) discarding as many non-polyp edges as possible and (2) determining on which side of the retained edges the polyps are present. As shown in Figure 2.7, our classification scheme analyzes a pair of oriented patches around each edge pixel and then depending on the appearance of the image pair classifies the underlying edge into either the polyp or non-polyp category and in the case of a polyp edge, identifies on which side of the boundary the polyp is

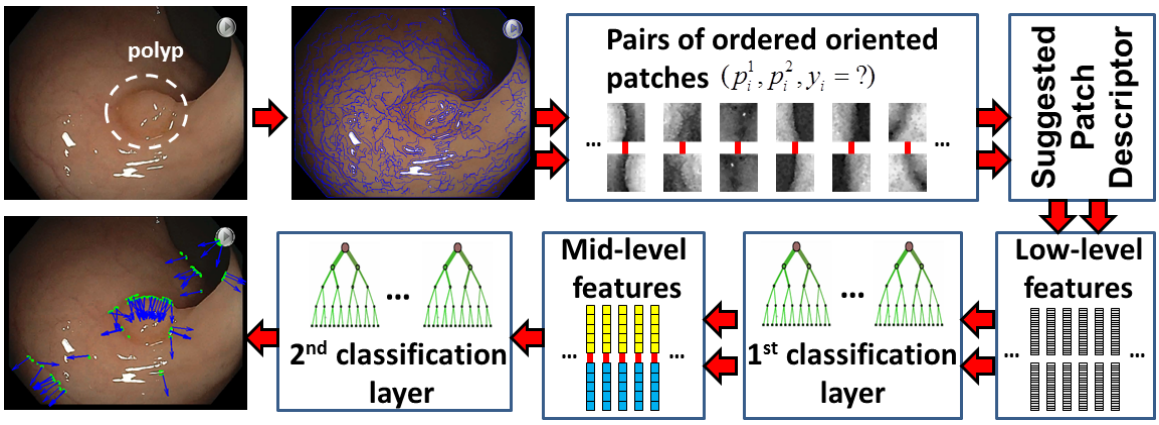


Figure 2.7: The test stage of the suggested classification scheme. Pairs of oriented patches extracted around each edge pixel are fed to the suggested feature extraction and classification schemes. In the end result, the green pixels indicate the edges that have passed the classification stage and the blue arrows indicate the possible location of a polyp for the retained edges. For more details, see Steps 1 and 2 of the pseudocode shown in Appendix A.

present. In the following, we first explain how the image pairs are collected and then describe the suggested classification scheme.

After ball tensor voting, each edge pixel will be assigned a structure tensor whose dominant Eigen vector $\vec{e}$ indicates the gradient orientation. However, since the gradient orientation has no particular directions, as shown in Figure 2.8, one can assume two normal directions for a given edge pixel, $\{n_i^1 = \vec{e}_i, n_i^2 = -\vec{e}_i\}$. The image is then interpolated along the normal and edge directions at each edge location, resulting in a pair of oriented patches $\{p_i^1, p_i^2\}$ given the two possible normal directions. Our classification scheme operates on each pair of patches and then combines their information not only to classify the underlying edge into polyp and non-polyp categories, but also to determine which normal direction among n_i^1 and n_i^2 points towards the polyp location. We refer to the determined normal direction as "voting direction" in

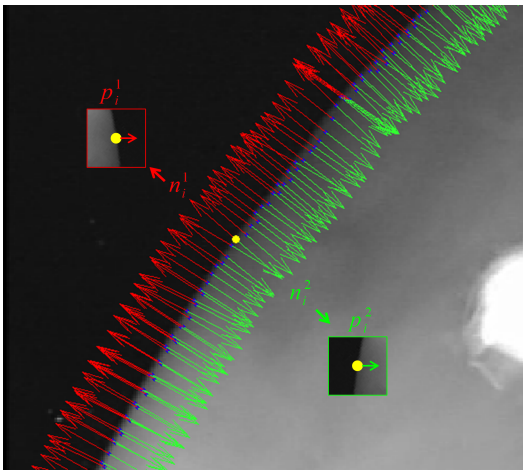


Figure 2.8: Pair of patches at each edge location. The green and red arrows show edge normal directions for a subset of edge pixels on a boundary. As seen, the angle between each pair of normal vectors is 180°, $\angle n_i^2 = \angle n_i^1 + \pi$. At each edge pixel, we extract two oriented patches according to the two normal directions (n_i^1, n_i^2). Each pair of patches (p_i^1, p_i^2) are horizontally mirrored.

the rest of this dissertation.

Our classification scheme has two stages. In the first stage, we learn the image appearance around the boundaries of the structures of interest in contrast to random structures in colonoscopy images. The structures of interest are chosen through a prior misclassification analysis and consist of polyps, vessels, lumen areas, and specular reflections. We train a five-class classifier based on the features generated by our patch descriptor to distinguish between such boundaries in complex endoluminal scenes. Thus, the classifier learned in the first stage measures the similarities between the input patches and the predefined structures by learning a non-linear metric in the low level feature space. The first layer can be also viewed as a feature enhancer that takes low-level image features from the proposed patch descriptor and produces mid-level similarity features.

The key to train the above classifier is to have a consistent presentation for each of the structures of interest. A consistent presentation is defined as a set of patches that all have the structure of interest on one side (e.g., right) and the background on the other side (e.g., left). For this purpose, one must choose the normal directions prior to patch extraction such that they all point towards or away from the structures of interest. Choosing normal directions in an arbitrary manner results in the arbitrary appearance of the structure of interest on the left or right side of the collected patches. In Figure 2.8, we show how the choice of the normal direction places the gray region in the left and right side of the resulting image patches.

To achieve patch consistency, we use the ground truth that we have created for the structures of interest. The ground truth definition depends on the structure of interest. For polyps, the ground truth is a binary image where the polyp region is white and the background is black. When extracting polyp patches, we choose the normal direction such that it points towards the white region. For lumen areas and specular spots, the ground truth is still a binary image but the white region corresponds to the dark lumen and the bright specular reflection, respectively. For vessels, the binary ground truth shows only the locations of vessels—we do not assume any preferred normal directions because the image appearance around the vessels is most often symmetric. For random structures, we collect image patches at random edge locations in colonoscopy images with arbitrary choice of normal directions.

The goal of the second classifier is to group the underlying edges into polyp and non-polyp categories and determine the desired normal directions. For this purpose, we train the second classifier based on the pairs of patches in the mid-level feature space, which is generated by the first classifier. We use pairs of patches because for a new image no information about the polyp location or about the desired normal directions is available. The classifier learns the desired normal directions by com-

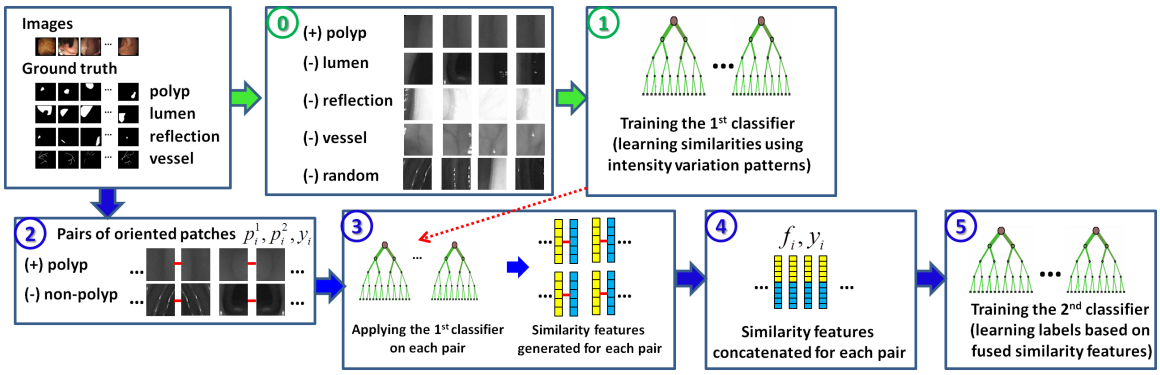


Figure 2.9: The training stage of the suggested classification scheme.

bining the information from each image pair. The training process of the suggested classification scheme is summarized as a pseudocode in Appendix B, is illustrated in Figure 2.9, and is further explained as follows:

- Step 0: We collect a stratified set of N_1 oriented patches around the boundaries of polyps, vessels, lumen areas, specular reflections, and random structures in the training images. Mathematically,

$$S^1 = \{(p_i^*, y_i) | y_i \in \{p, v, l, s, r\}, i = 1, 2, ..., N_1\}.$$

  Note that asterisk indicates that the patches are extracted according to the desired normal directions, which are available given the ground truth for the training images.

- Step 1: Once patches are extracted, we train a five-class classifier in the low level feature space created by the proposed patch descriptor. The output of this classifier is an array of probabilities for the five object classes. Compared with the low level input features, which encode local intensity variations, the generated output array contains mid-level features that measure the global similarity between the underlying input patches and the boundaries of the predefined structures.

- Step 2: We collect a stratified set of N_2 pairs of oriented patches from polyp and non-polyp boundaries in the training images. We order the pair of patches $\{p_i^1, p_i^2\}$ around the i^{th} edge according to the angles of their corresponding normal vectors, $\angle n_i^1 < \angle n_i^2$. This convention is to keep patches in a consistent order. We assign a label $y_i \in \{0, 1, 2\}$ to each pair of patches, where "0" is for a non-polyp edge, "1" is for a polyp edge with n_i^1 indicating the desired normal direction, and "2" is for a polyp edge with n_i^2 indicating the desired normal direction. Mathematically,

$$S^2 = \{(p_i^1, p_i^2, y_i) | y_i \in \{0, 1, 2\}, i = 1, 2, ..., N_2\}.$$

- Step 3: We extract low level features from each pair of patches using the suggested patch descriptor and then apply the classifier trained in the first layer, resulting in two arrays of mid-level features.

- Step 4: For each pair of oriented patches, the corresponding array of mid level features are concatenated to form a feature vector, $\{(f_i, y_i) | y_i \in \{0, 1, 2\}, i = 1, 2, ..., N_2\}$.

- Step 5: Once all feature vectors are collected, we train a three-class classifier to learn both edge labels and edge normal directions (embedded in y_i).

In Figure 2.7, we show the test stage of the suggested classification scheme. Given a test image, we first obtain the corresponding edge map and then extract pairs of oriented patches around each detected edge. Next, we compute the low level features for each pair of patches using the suggested patch descriptor. The low level features are then fed to the first classification layer, resulting in mid level features that are further concatenated for each pair of patches. The second classification layer is then

applied to the concatenated mid-level features, producing three probabilistic values corresponding to three possible labels for each edge pixel. The underlying edge of each image pair is declared as a polyp edge if the polyp probability dominates that of the other two classes:

$$p(y_i = c) > p(y_i \neq c),\ c \in \{1, 2\},$$

which is met if and only if

$$p(y_i = c) > 0.5,\ c \in \{1, 2\}.$$

Therefore, the underlying edge of each pair of patches is classified according to the following rule:

$$\begin{cases} \text{“polyp” and } n_i^* \leftarrow n_i^1 & \text{if } p(y^i = 1) > 0.5 \\ \text{“polyp” and } n_i^* \leftarrow n_i^2 & \text{if } p(y^i = 2) > 0.5 \\ \text{“non-polyp”} & \text{otherwise,} \end{cases} \tag{2.2}$$

where n_i^* is the desired normal direction or voting direction. The other alternative to Eq. 2.2 is to assign the edge pixel to the class with maximum probability, but this cannot handle uncertain situations where the probability associated with one class is only slightly larger than each of the other two individual classes. Once all the edges are classified, non-polyp edges are removed from the edge map and the rest along with their corresponding voting directions are sent to the vote accumulation scheme for polyp localization.

We conclude this section with a discussion on the necessity of the second classification stage. Consider the edge classification scenario where we use only the first classifier for edge classification. After a pair of patches passes the first classifier, we obtain two sets of probabilities (mid-level features). To determine a polyp edge, one can compare the polyp probabilities between the two patches and check whether the

larger probability dominates the rest. The desired normal direction is also determined as the normal direction of the patch with the larger probability. However, a problem arises when there exist more than one dominating class. For instance, consider the following two sets of probabilities obtained for a pair of patches, $\{\mathbf{0.7}, 0.1, 0.0, 0.0, 0.2\}$ and $\{0.0, 0.3, 0.1, \mathbf{0.6}, 0.0\}$, which suggests two dominating classes. The first set corresponds to patch p_i^1 around i^{th} edge resembling the appearance of a polyp boundary while the second set corresponds to patch p_i^2 resembling the boundary appearance of specular reflections. This produces uncertainty as to the decision regarding the underlying edge pixel. The choice is to either rely on the first patch and declare a polyp edge with edge normal being n_i^1 or consider information from the counterpart patch and declare a non-polyp edge. One way to resolve the issue is to declare a polyp edge if there exists only one dominating class; however, this may result in a sub-optimal solution. To eliminate the need for such user-defined rules, we train a second classifier in the mid-level feature space to learn such relationships and utilize them for more accurate edge classification.

2.2.1.4 Voting Scheme.

Our voting scheme is designed to localize polyps by identifying their curvy heads. This is achieved by generating a heat map where a higher temperature is assigned to the regions that are surrounded by curved boundaries.

The voters that participate in our voting scheme are the the edges that have passed the classification stage. Each voter has a voting direction n_i^* and a classification confidence $C_{v_i} = max(p(y^i = 1), p(y^i = 2))$. Our voting scheme begins with grouping the voters into K categories according to their voting directions, $V^k = \{v_i | \frac{k\pi}{K} < mod(\angle n_i^*, \pi) < \frac{(k+1)\pi}{K}\}$, $k = 0...K-1$. The voters in each category start casting votes at their surrounding pixels according to their voting directions and classifica-

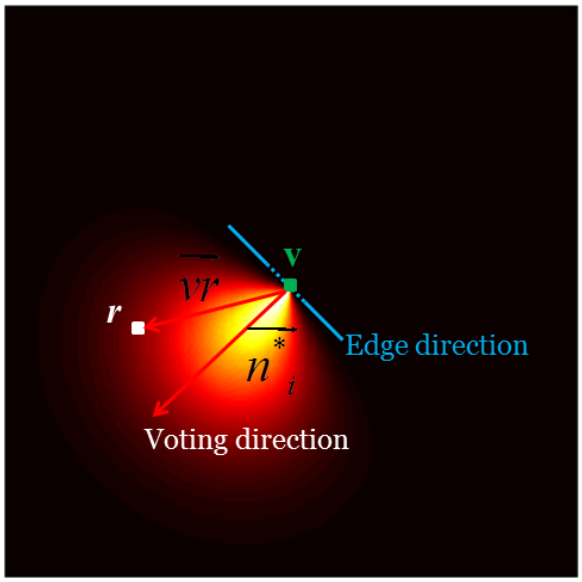


Figure 2.10: The generated voting map for an edge pixel lying at 135 degree.

tion confidence. The vote cast by the voter v at a receiver pixel $r = [x, y]$ is computed as

$$M_v(x,y) = \begin{cases} C_v \exp(\frac{-\|\vec{vr}\|^2}{\sigma_F}) \cos(\angle \vec{n^*} \vec{vr}), & \text{if } \angle \vec{n^*} \vec{vr} < \pi/2 \\ 0, & \text{if } \angle \vec{n^*} \vec{vr} \geq \pi/2 \end{cases} \tag{2.3}$$

where the exponential and cosinusoidal functions enable smooth vote propagation, which we will later use to determine the likelihood of a polyp candidate. In Eq. 2.3, $\vec{vr}$ is the vector connecting the voter and receiver, σ_F controls the size of the voting field, and $\angle \vec{n^*} \vec{vr}$ is the angle between the voting direction $\vec{n^*}$ and $\vec{vr}$. In Figure 2.10, we show the voting field for an edge pixel lying at 135 degree. As seen, the votes are cast only in the region pointed by the voting direction. Such selectivity arises from the condition set on $\angle \vec{n^*} \vec{vr}$, which prevents the voters from casting votes in the opposite direction.

Once all the voters in each category cast votes at their neighboring pixels, an accumulator adds up the votes received at each pixel and generates a voting map (see Figure 2.11). The resultant K voting maps are then multiplied and the pixel

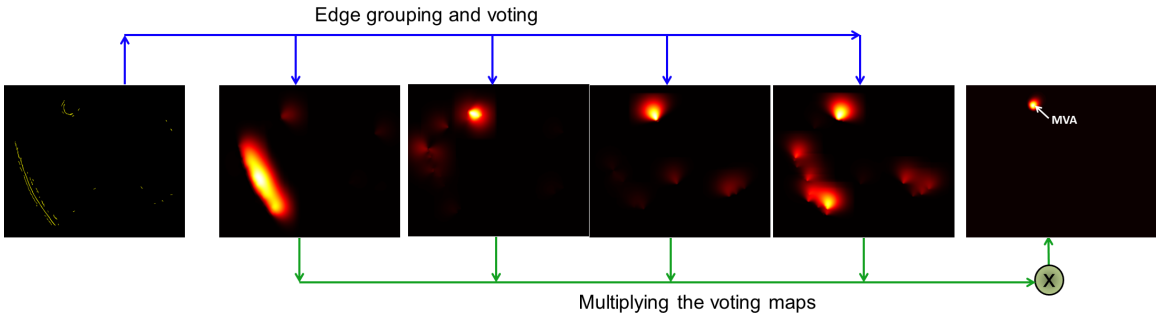


Figure 2.11: Voting process. Given a refined edge map, the voting scheme begins with grouping the edges into $K = 4$ categories. The votes cast in each category are then accumulated in a separate 2D array, resulting in 4 voting maps whose multiplication results in the final voting map. The point with maximum vote accumulation (MVA) is considered to be a polyp candidate.

with maximum vote accumulation (MVA) is considered as the location of the polyp candidate. Mathematically,

$$MVA = \arg\max_{x,y} \prod_{k=0}^{K-1} \sum_{v \in V^k} M_v(x, y). \tag{2.4}$$

It is essential for our voting scheme to prevent vote accumulation in the regions that are surrounded by low curvature boundaries, because such regions in general cannot present the curvy heads of polyps. This was achieved in our voting scheme by grouping edges prior to vote casting and multiplying the resultant voting maps. The rationale is that pixels on low curvature boundaries contribute to only a small fraction of the K to-be-multiplied voting maps. To clarify this, we generate a synthetic image containing a polyp-like structure and a set of parallel lines, and then compare the resulting voting maps with and without the map multiplication strategy. In Figure 2.12(a), we show the vote accumulation map when the votes cast by the voters all accumulated in one voting map, $\sum_v M_v(x, y)$. As seen, the votes are accumulated in two regions: inside the curvy structure which is desirable, and between the parallel

lines which is undesirable. In Figure 2.12(b), we show the voting map for the same image when edge grouping and map multiplication are employed. As seen, the accumulator assigns low values to the region between the parallel lines, and high values to the region inside the polyp-like structure.

Another important characteristic of our voting scheme is the utilization of voting directions that, as shown in Figure 2.10, limits a voter to cast votes only along its assigned voting direction. Ignoring voting directions (n_i^*) and allowing voters to vote along both possible normal directions (n_i^1, n_i^2) result in vote accumulation on both sides of the boundaries, which often leads to polyp mislocalization. This is illustrated in Figure 2.12(c) where no selectivity in voting directions causes polyp mislocalization, but including such a selectivity allows for correct polyp localization (Figure 2.12(d)). Our voting scheme is summarized in Step 3 of the pseudocode shown in Appendix A.

2.2.2 Stage 2: Candidate Classification

Our candidate classification method begins with estimating a bounding box around each polyp candidate. It then proceeds with a novel score fusion framework based on convolutional neural networks (CNNs) Krizhevsky *et al.* (2012) where the goal is to assign a confidence value to each generated candidate. Candidates with higher confidence are more likely to be a polyp.

2.2.2.1 Bounding Box Estimation

To measure the extent of the polyp region, we estimate a narrow band around each candidate, so that it contains the voters that have contributed to vote accumulation at the candidate location. Thus, the desired narrow band will enclose the polyp boundary and thus can be used to estimate a bounding box around the candidate location. As shown in Figure 2.13(a), the narrow band B consists of a set of radial

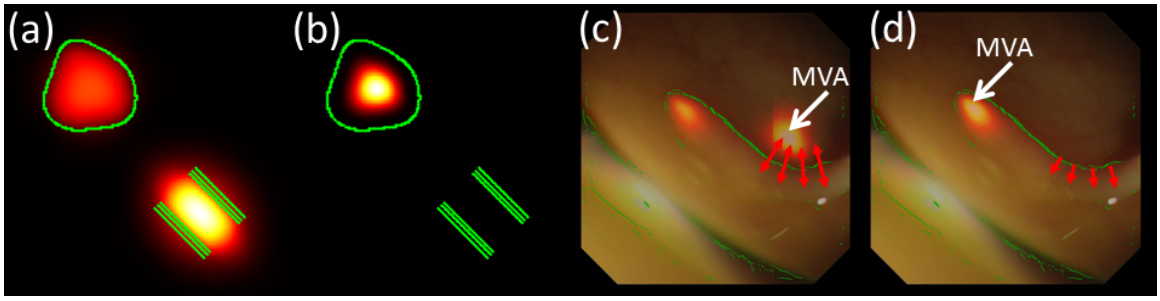


Figure 2.12: Using voting directions and map multiplication is essential for correct polyp localization. (a) The resultant voting map for a synthetic scene when the generated votes are accumulated in one voting map. Undesirably, large vote accumulation is observed around the parallel lines. (b) The resultant voting map for the same scene when the generated votes are accumulated in $K = 4$ voting maps and then multiplied (see Eq. 2.4). As seen, votes are desirably accumulated in the object with the curvy boundary. (c) A colonoscopy image and its corresponding voting map when the voters cast votes along both possible normal directions. As seen, the polyp candidate (MVA) is placed outside the polyp region. (d) The same colonoscopy image and its corresponding voting map when the voters cast votes only along the inferred voting directions. The polyp has been localized successfully. Therefore, incorporating voting directions improves polyp localization. We highlight in green the voters or the edges retained after the classification scheme.

lines ℓ_θ parameterized as $\ell_\theta : MVA + t[\cos(\theta), \sin(\theta)]^T, t \in [t_\theta - \frac{\delta}{2}, t_\theta + \frac{\delta}{2}]$, where δ is the bandwidth, and t_θ is the distance between the candidate location and the corresponding point on the band skeleton at angle θ. To form the band around a polyp candidate (MVA), one needs to determine the bandwidth δ and a set of distances t_θ. Once the band is formed, the bounding box is localized so that it fully contains the narrow band around the candidate location (see Figure 2.13(a)).

We introduce the isocontours of a voting map as a means to estimate the unknown

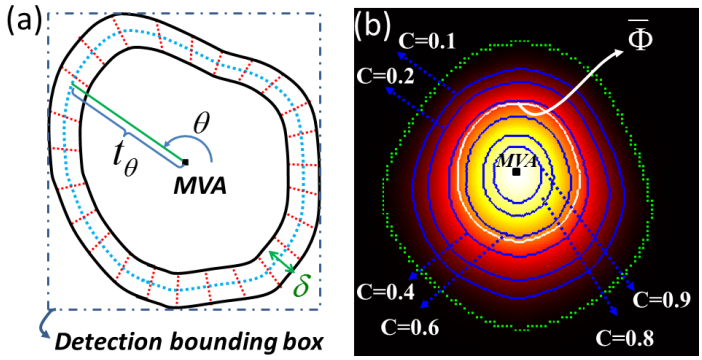


Figure 2.13: Determining a narrow band around a polyp candidate. The narrow band consists of the two black contours and the enclosed area. We use the narrow band to assign a probabilistic score to each polyp candidate. The blue curve is the band skeleton and the red lines are a subset of the radial line segments that represent the band in its discrete form.

parameters of the corresponding narrow band. The isocontour Φ_c of the voting map V is defined as $\Phi_c = \{(x, y)|V(x, y) = c \times M\}$ where M denotes the maximum of the voting map and c is a constant between 0 and 1. As shown in Figure 2.13(b), the isocontours of the voting map, particularly those located farther away from the candidate, have the desirable feature of following the shape of the actual boundary from which the votes have been cast at the candidate location. Therefore, one can estimate the narrow band's parameters from the isocontours such that the band encloses the object's boundary. However, in practice, the shape of far isocontours are undesirably influenced by other nearby voters in the scene. We therefore obtain the representative isocontour $\bar{\Phi}$ by computing the median shape of the isocontours of the voting map (see Figure 2.13(b)). We have experimented with different sets of isocontours and found out that as long as their parameter c is uniformly selected between 0 and 1, the resulting representative isocontour serves the desired purpose.

Let d^i_{iso} denotes the distance between the i^{th} point on the representative isocoun-

tour $\bar{\Phi}$ and the candidate location. We use d^i_{iso} to predict d^i_{obj}, the distance between the corresponding point on the object boundary and the candidate location. For this purpose, we employ a second order polynomial regression model

$$d^i_{obj} = b_0 + b_1(d^i_{iso}) + b_2(d^i_{iso})^2, \tag{2.5}$$

where b_0, b_1, and b_2 are the regression coefficients that are estimated using a least square approach. Once the model is constructed, we take the output of the model d_{obj} at angle θ with respect to MVA as t_θ and the corresponding prediction interval as the bandwidth δ.

2.2.2.2 Probability Assignment

We propose a score fusion framework based on convolutional neural networks (CNNs) that can learn and integrate color, texture, shape, and temporal information of polyps in multiple scales for more accurate candidate classification. To the best of our knowledge, CNNs have never been explored for detecting polyps in colonoscopy videos. The latest generation of CNNs has proven to be very successful in object detection and recognition tasks, significantly outperforming the best performance reported for many databases Ciresan *et al.* (2012). The attractive feature of CNNs is that they jointly learn a multi-scale set of image features and a discriminative classifier during a supervised training process. While CNNs are known to learn discriminate patterns from raw pixel values, it turns out that preprocessing and careful selection of the input patches can have a significant impact on the performance of the subsequent CNNs. Specifically, we have found out that partial illumination invariance achieved by histogram equalizing the input patches significantly improves the performance of the subsequent CNNs and that curse of dimensionality caused by patches with more than three channels results in CNNs with inferior performance.

Considering these observations, we propose a three-way image presentation that is motivated by the three major types of polyp features suggested in the literature, which are (1) color and texture clues, (2) temporal features, and (3) shape in context. Let I_t^{rgb} denote a color colonoscopy image captured at time t, (C_x, C_y) denote a polyp candidate, and w denote the estimated width of the square bounding box measured in pixels. We further use I_t^e and I_t^v to present the resulting refined edge map and voting map for the image captured at time t. Given these notations, we define the following three sets of patches:

- P_C consists of three-channel color patches that represent color and texture information around polyp candidates. For a polyp candidate C detected in I_t^{rgb}, we extract a $w \times w \times 3$ patch from the histogram-equalized version of I_t^{rgb}. Our experiments show that partial illumination invariance, achieved by histogram equalization, significantly improves the accuracy of the subsequent CNN.

- P_S consists of three-channel patches that represent shape in context around polyp candidates. For a polyp candidate C detected in I_t^{rgb}, we extract a $w \times w \times 3$ patch, where the first channel is selected from the gray scale version of I_t^{rgb}, the second channel from the refined edge map I_t^e, and third channel from the voting map I_t^v. Our refined edge map is preferred over the complete edge map because the latter contains a large amount of spurious edges associated with specular spots (particularly on polyp surfaces) that can hinder the learning of the essential shape information of polyps.

- P_T consists of three-channel patches that represent temporal information around polyp candidates. For a polyp candidate C detected in I_t^{rgb}, we extract a $w \times w \times 3$ patch, where the i^{th} channel is selected from the histogram-equalized gray scale version of I_{t-i-1}^{rgb}. Unlike Park *et al.* (2012), we do not include information

from the future frames because it is not a realistic assumption about how a real-time polyp detection system should operate.

The above-explained sets of patches have one limitation: they do not reflect all possible variations in scale and orientation of polyps. To overcome this limitation, we enrich each set by extracting multiple patches at each candidate location. Specifically, given a polyp candidate C and the corresponding $w \times w$ bounding box, we extract patches at N_t locations $C + (\Delta_x^{(t)}, \Delta_y^{(t)})$ and N_s scales $\alpha^{(s)} \times w$. To obtain a rotation-invariant presentation, we also include N_r rotated versions of the collected patches. This results in $N_r \times (N_t + 1) \times N_s$ patches extracted around each polyp candidate.

Once the enriched set of patches are generated, we label each individual patch depending on whether the underlying candidate is a true or false positive. Such a labeling is possible because the ground truth is available for the training datasets. We then use the labeled enriched datasets for training CNNs. Specifically, we train a CNN for each collected dataset, resulting in three CNNs that will be used when a test video is available. During the test stage, each frame of the video will undergo the candidate generation stage, resulting in a polyp candidate for the given frame. The candidate classification begins with extracting the three sets of patches (P_C, P_T, P_S) around the polyp candidate. Next, each set of patches is sent to the corresponding CNN, which has previously been trained. Each CNN analyzes the input set of patches, generates confidence values for all the patches within the set, and then computes the maximum confidence value for the input set of patches. This process is repeated for all the three CNNs, resulting in three confidence values. The final probability of being a polyp is computed as the average of the resulting confidence values. In Figure 2.14, we show the test stage of the suggested score fusion framework.

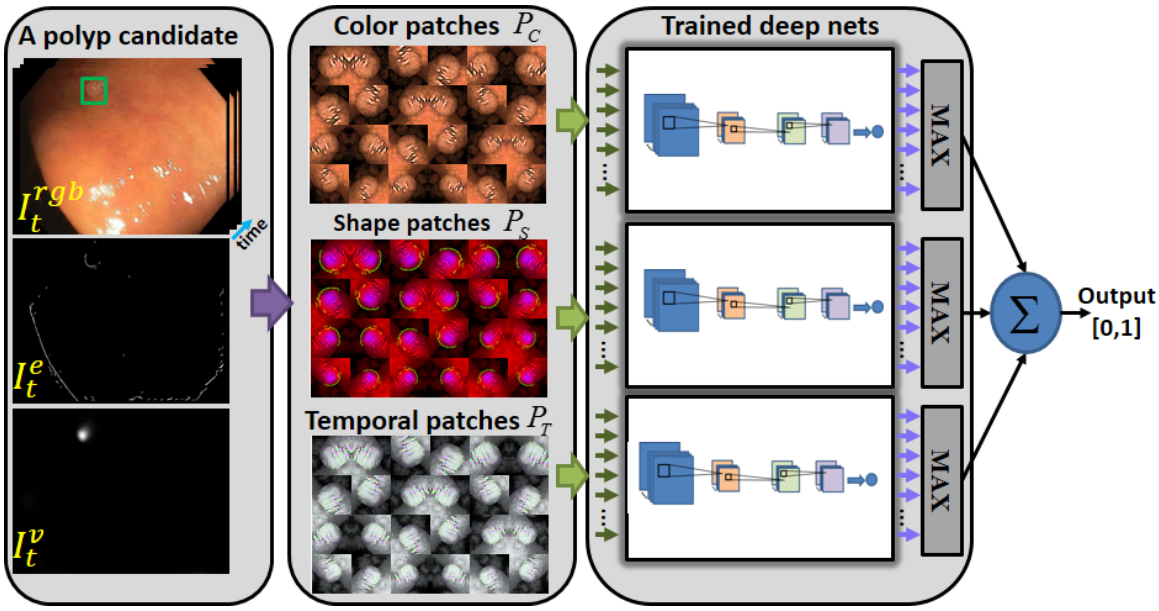


Figure 2.14: The suggested score fusion framework based on convolutional neural networks.

2.3 Data

We use a publicly available polyp database *CVC-ColonDB* Bernal *et al.* (2012) and our collection of colonoscopy videos called ASU-Mayo Clinic polyp database for evaluation.

2.3.1 CVC-ColonDB

CVC-ColonDB Bernal *et al.* (2012) is a publicly available polyp database collected and de-identified at the University of Barcelona, Spain. This database contains 300 colonoscopy images selected from 15 short colonoscopy videos so that maximum variation in scale and view angles of the polyps are captured. Each image in this database contains a polyp. The ground truth corresponding to each colonoscopy image is a binary (black and white) image where the polyp and background regions are indicated in white and black, respectively. This database is available at

http://mv.cvc.uab.es/projects/colon-qa/cvccolondb.

2.3.2 ASU-Mayo Clinic Polyp Database

The ASU-Mayo Clinic polyp database is the first, largest, and a constantly growing set of short and long colonoscopy videos, collected and de-identified at the Department of Gastroenterology at Mayo Clinic in Arizona. The videos are selected so as to display maximum variation in colonoscopy procedures: some videos are high resolution, while some are recorded in lower resolution; some videos display a careful colon examination, while some show a hasty colon inspection; some videos show a well-prepared colon, while some show a large amount of fecal content. In addition, some videos have biopsy instruments in them, while others have a play logo at the top-right of the frames. Our database of colonoscopy videos is available at www.polyp2015.com.

Currently, our database consists of 40 short colonoscopy videos, of which half are positive and half are negative shots. We define a positive shot as a sequence of frames that shows a unique polyp from different view angles. A negative shot is a segment of colonoscopy video that does not contain a polyp. We have randomly halved the database at video level into the training set containing 3,800 frames with polyps and 15,100 frames without polyps, and the test set containing 5,700 frames with polyps and 13,200 frames without polyps. Each frame in this database comes with a ground truth image or a binary mask that indicates the polyp region. If a frame contains no polyp, the corresponding ground truth image will be completely black. The ground truth images for the collected videos have been created by volunteer students at Arizona State University, and have been reviewed and corrected by a trained expert. In Figure 2.15, we show examples of polyps from our database along with their corresponding ground truth images.

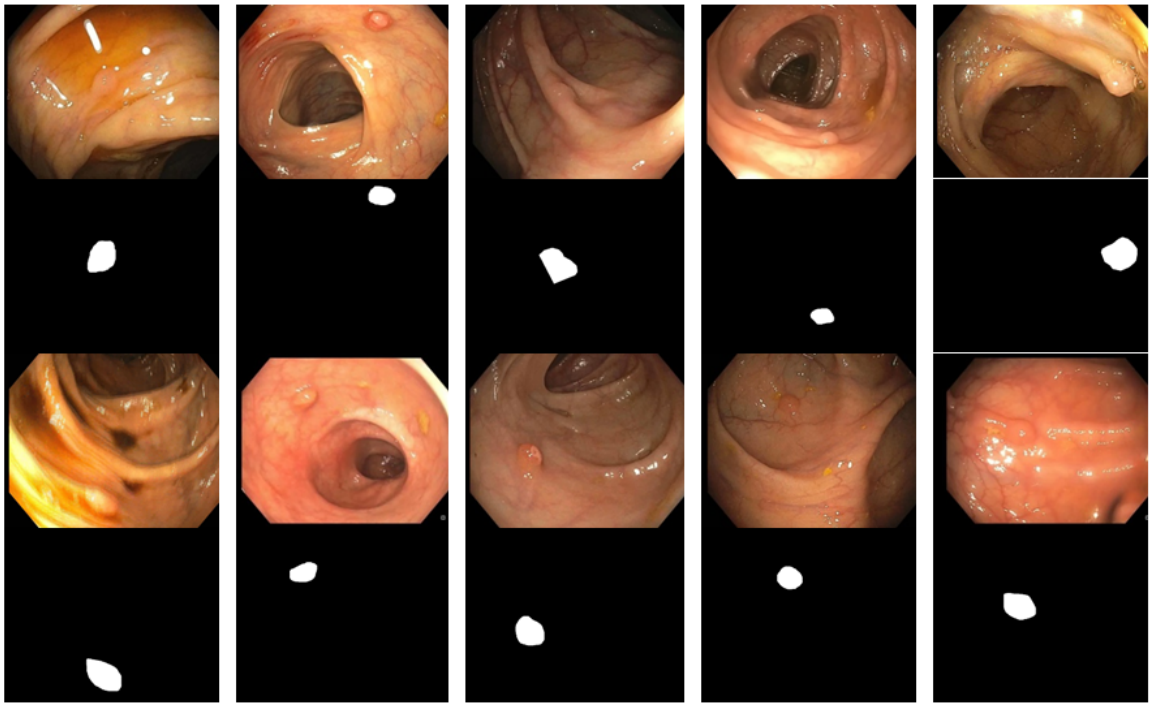


Figure 2.15: Examples of polyps and the corresponding ground truth images from the ASU-Mayo Clinic Database.

2.4 Experiments

We have randomly split the ASU-Mayo Clinic database at video level into training and test sets. We have used the training set to adjust the parameters of the suggested system and train the learning-based modules, and used the test set for evaluation. In the following, we first explain the training process and parameter tuning of the suggested polyp detection system and then present the results on the test set.

2.4.1 Model Tuning

Edge detection. We use Canny edge detector to obtain initial edge maps of colonoscopy images. The key parameter of Canny edge detector is the degree of Gaussian smoothing σ_g, which is essential to removing fake and spurious edges. A

low degree of smoothing results in over-cluttered edge maps with a large number of edges associated with image noise. We experimentally found out that over-cluttered edge maps degrade the performance of the subsequent steps of the suggested polyp detection system. On the other hand, a high degree of smoothing results in aggressive removal of the edges, generating edge maps in which the polyp boundaries are missing. Our experiments using the training dataset indicate that $\sigma_g = 3$ yields a good trade-off between detecting polyp edges and removing noisy edges.

Feature Extraction. Our feature evaluation method has three parameters: width of sub-patches w, height of sub-patches h, and the number of the DCT coefficients selected from each sub-patch n. To tune these parameters, we trained a number of classifiers and investigated what configuration of the parameters achieves the highest classification performance. For this purpose, we collected 50,000 oriented patches around polyps and other boundaries in colonoscopy images. We then randomly split this set of patches into two subsets: one for training a random forest classifier and the other for testing. According to our experiments, employing 8×16 sub-patches and selecting $n = 3$ DCT coefficients from each sub-patch achieves the best classification performance.

To evaluate the competence of the suggested feature extraction, we have used ROC analysis to compare the performance of our method with that of other widely-used descriptors, such as HoG Dalal and Triggs (2005), LBP Ojala *et al.* (2002), and Daisy Tola *et al.* (2010). For fair comparisons, we used the publicly available implementations of HoG [1] , LBP [2] and Daisy [3] , tuned their parameters using the dataset that we employed for tuning the parameter of the suggested feature extraction,

[1] lear.inrialpes.fr/pubs/2005/DT05/

[2] www.cse.oulu.fi/CMV/Downloads/LBPMatlab

[3] cvlab.epfl.ch/software/daisy

and then performed evaluation using a common test set. For LBP, we divided each patch into cells of size 8x8 and for each cell we computed the normalized histogram of rotation invariant uniform patterns using a 3x3 neighborhood around the pixels. The resulting 10-bin histograms from all the cells were then concatenated to form the final feature vector. For HoG, we used cells of size 8x8 pixels and blocks of size 2x2 cells or 16x16 pixels. We computed a gradient histogram with 9 orientation bins for each block and then concatenated the resulting histograms. For Daisy, we defined three concentric rings around the center of the patch and then selected 8 equally spaced points on each ring. Next, we concatenated 8-bin gradient histograms computed at each of the selected points.

In Figure 2.16, we compare the ROC curve of the suggested feature extraction method with that of the other competing methods. As seen, our descriptor surpasses HOG and LBP with a large margin and outperforms Daisy with a smaller yet statistically significant margin [4] ($p < .0001$). In addition to superior classification performance, our descriptor runs approximately 30 times faster than its closest competitor (Daisy), which makes it further amenable to the suggested classification scheme.

To further analyze our patch descriptor, we visualize the variable importance computed by the random forest classifier for each of the extracted features. Random forest calculates the importance of feature f_i in each tree and then takes their average to compute the overall importance of feature f_i. To measure importance of feature f_i in each tree, random forest permutes the values of this feature in the out-of-bag samples randomly and then measures the subsequent decrease in the classification performance. A feature is considered important if the corresponding permutations

[4] DeLong's method using MedCalc for Windows, version 13.3 (MedCalc Software, Ostend, Belgium)

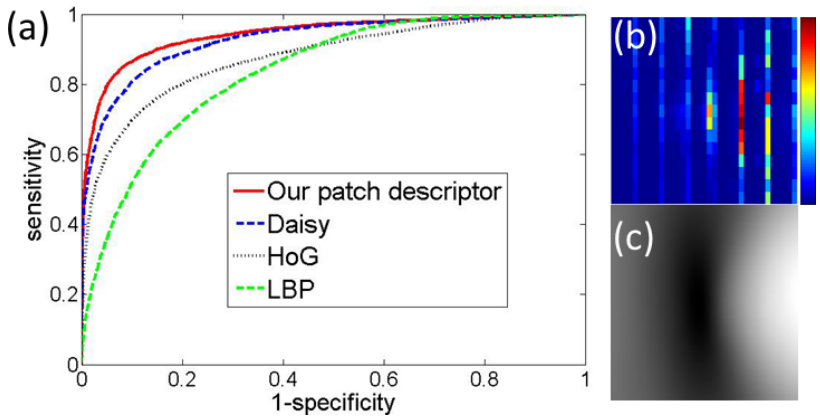


Figure 2.16: Evaluation of our patch descriptor. (a) Our patch descriptor surpasses HOG and LBP with a large margin and outperforms Daisy with a smaller yet statistically significant margin. Our descriptor runs approximately 30 times faster than Daisy. (b) Feature importance computed by the random forest classifier that is trained using the suggested descriptor. For visualization purposes, feature importance values are reshaped into a matrix with the same size as the average image patch. Comparing the importance map with (c) average image appearance around polyp boundaries suggests that while the discriminative patterns are found all over the patches, they are more densely located inside the polyp region and across the polyp boundary.

minimally affect classification performance. We collect the variable importance of all the 315 features and then reshape them into a matrix form such that each feature gets mapped to the part from which it has been extracted. To visually compare the importance map with the average of polyp patches (shown in Figure 2.16(c)), we scale up this matrix to the same size as the input image patches, as shown in Figure 2.16(b). As seen, while the discriminative patterns are found all over the patches, they are more densely located inside the polyp region and across the polyp boundary. The relatively less number of important features on the background side (left side) of the patches can be explained by the large variability in the backgrounds of polyps.

Edge classification. To train our 2-stage edge classifier, We collected $N_1 = 100,000$ oriented image patches with approximately 20,000 samples for each of the five chosen structures to train a random forest classifier in the first stage, and $N_2 = 100,000$ pairs of oriented image patches to train another random forest classifier in the second stage. The choice of a random forest classifier is motivated by its recent success in a variety of computer vision and medical image analysis applications where it outperformed other widely-used classifiers such as AdaBoost and support vector machines Criminisi and Shotton (2013). The two main ingredients of a random forest classifier are bagging of a large number of fully grown decision trees and random feature selection at each node while training the trees, which together achieve low generalization error and high quality probabilistic outputs. In our experiments, we kept adding decision trees to the random forest classifiers until the decreasing trend of out-of-bag error converged. According to our experiments, 100 fully grown decision trees achieved a stable out-of-bag error for both random forest classifiers.

Voting scheme. The suggested voting scheme has two parameters: the size of voting fields σ_F and the number of voting categories K. However, one cannot tune these two parameters by only considering the performance of the voting scheme. This is because the voting scheme also affects the performance of the subsequent stage that is bounding box estimation. To clarify this relationship, recall that bounding box estimation is based on a regression model (see Eq. 2.5) whose predictor variable is a function of the voting scheme. As a result, one can obtain different regression models for different configurations of the voting scheme. Therefore, we select the optimal choice of the parameters so as to optimize both the voting method and the bounding box estimation method. For this purpose, we base our study on a shape generator model that allowed us to to generate a large number of synthetic polyp-like objects. Using the actual polyps may limit the generalizability of this parameter study as

the existing polyps in the database may not completely reflect the significant shape variation among polyps.

We used the synthetic shapes generated by our model to construct an initial set of regression functions for different values of K and σ_F. To find the best model, we first construct an initial set of regression functions for different values of K and σ_F. We then narrow down the initial set of regression models by finding the optimal value of K and σ_F. More specifically, we take the following steps:

- First, we generate 3,000 objects at three different scales corresponding to small, medium, and large polyps. To do so, we use our shape generator model and choose $\mu_r \in \{20, 40, 60\}$, and set $\sigma_r = 0.2\mu_r$, $\mu_a = 1$, and $\sigma_a = 0.1$.

- Second, we perform the voting scheme for each generated object using $K \in \{2, 3, 4, 5, 6\}$ and $\sigma_F \in \{60, 70, 80, 90, 100\}$. For each object, the set of voters consists of all the edge pixels that form the object contour. To initiate the voting process, each voter must be assigned a voting direction. We first obtain the edge direction for an edge pixel using ball tensor voting and then determine its voting direction such that it points towards inside the corresponding object.

- Third, we find the representative isocontours of the generated voting maps, and collect pairs of (d^i_{obj}, d^i_{iso}) from the boundaries of the objects and the representative isocontours.

- Fourth, we build a regression model based on the collected pairs for each configuration of the parameters and then evaluate model fitness.

In Table 2.2, we show the R^2 coefficient for the constructed regression models. The coefficient of determination or R^2 indicates how well the data fit into a statistical model. The higher the R^2, the better the fit. If the model explains all the

Table 2.2: R^2 coefficient for the regression models constructed under different configurations of the voting scheme. Our experiments suggest that categorizing voters into four categories prior to vote casting achieves the highest boundary localization accuracy. We also observe that using larger voting fields yields higher accuracy for the synthetic objects. However, employment of large voting fields is not effective for real colonoscopy images due to undesirable effects of the voters that lie on the nearby structures. We will investigate the impact of on the overall polyp detection performance in Section 2.4.2.

		Size of the voting field				
		σ_F=60	σ_F= 70	σ_F=80	σ_F=90	σ_F=100
# voting groups	K=2	0.863	0.905	0.921	0.928	0.928
	K=3	0.830	0.890	0.916	0.930	0.936
	K=4	0.820	0.891	0.924	0.941	**0.951**
	K=5	0.812	0.889	0.923	0.941	0.950
	K=6	0.794	0.876	0.914	0.934	0.946

variation in data, R^2 will reach a maximum of 1. As seen, model fitness or the ability to localize object contours initially increases but then decreases as the number of voting categories increases. This is because, too few voting categories results in large angle quantization error. Thus, the shapes of the representative isocontours will not follow the shape of the object. On the other hand, too many quantization level favor pure circular objects, limiting the localization capability for objects that deviate from a complete circle. We choose $K = 4$ for the subsequent experiments in this article. Model fitness, however, monotonically increases as the size of the voting

fields increases. This is because, the parts of the boundaries that are farther away from the object centers need larger voting fields in order to contribute to vote accumulation in the centers of the objects. However, in practice, σ_F cannot be set to an arbitrary large value since it causes voting interference from the voters lying on the other nearby objects. Moreover, a large voting field incurs heavy computational burden. We will investigate the impact of σ_F on the overall polyp detection performance in Section 2.4.2.

Probability assignment. In the suggested system, a mixture of CNNs determine the probability of a candidate being a polyp. We used Krizhevsky's GPU implementation Krizhevsky *et al.* (2012) of CNNs for our experiments. To train CNNs, we collected 400,000 32x32 patches for P_C, P_T, and P_S where half of the patches were extracted around false positive candidates and the rest around true positive candidates. We have experimented with different CNN layout using the training data and observed relatively stable behavior among the trained CNNs. We have used the layout shown in Figure 2.17 for all the CNNs used in this paper. It is a common practice Krizhevsky *et al.* (2012) when training CNNs to randomly crop smaller patches from the original training patches. Such an image cropping mechanism achieves partial robustness against small image translation and also helps avoid over-fitting to the training data. Following this practice, our network begins with taking 24x24 sub-patches from random locations in the original 32x32 patches. The network then follows by two convolution blocks and two locally connected layers. Each convolution block is comprised of a convolution layer with rectified linear activation functions followed by a local normalization layer and a max pooling layer with a down-sampling factor of two. The first convolution layer learns 5x5 discriminative patterns from each channel of the training patches. The second convolution layer learns 5x5 filters as well but from the feature maps generated in the first convolution block. The last two

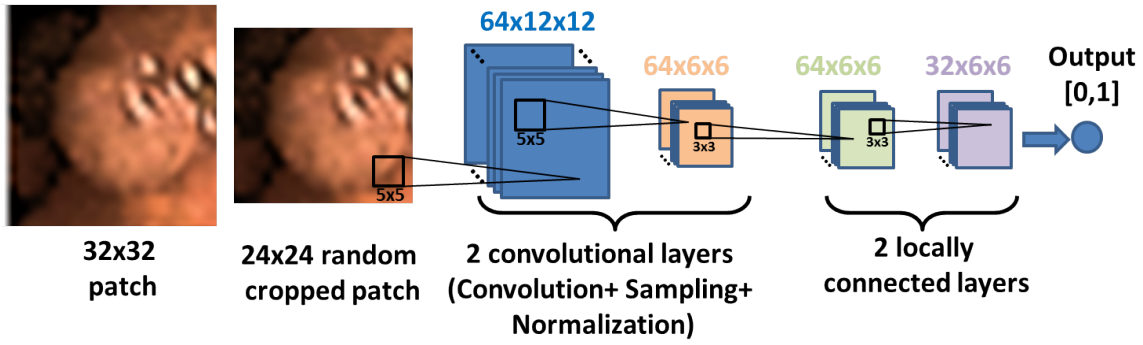


Figure 2.17: Network layout for the deep convolution networks. For illustration, the input patch is selected from P_C, but the same layout is used for the other types of training patches.

locally connected layers have no weight sharing and thus each learns 1,152 filters of size 3x3.

2.4.2 Model Evaluation

We evaluate the performance of the suggested polyp detection system using (1) the widely-used FROC analysis and (2) a new detection latency analysis that we specifically designed for colonoscopy. In the following, we first explain performance metrics in the context of colonoscopy and then present the analyses.

We define a true positive (TP) as a polyp candidate that falls inside the ground truth corresponding to a polyp. A false positive is defined as a polyp candidate that falls outside the ground truth associated with polyps. A false negative is a polyp instance that has not been detected by our polyp detection system. With these definitions, we compute sensitivity or recall as $\frac{TP}{TP+FN}$ and precision as $\frac{TP}{TP+FP}$.

FROC analysis. Free-response ROC (FROC) is a performance curve that is commonly used for evaluating the performance of computer-aided detection systems. This 2D plot shows the lesion localization rate with respect to the average number of false

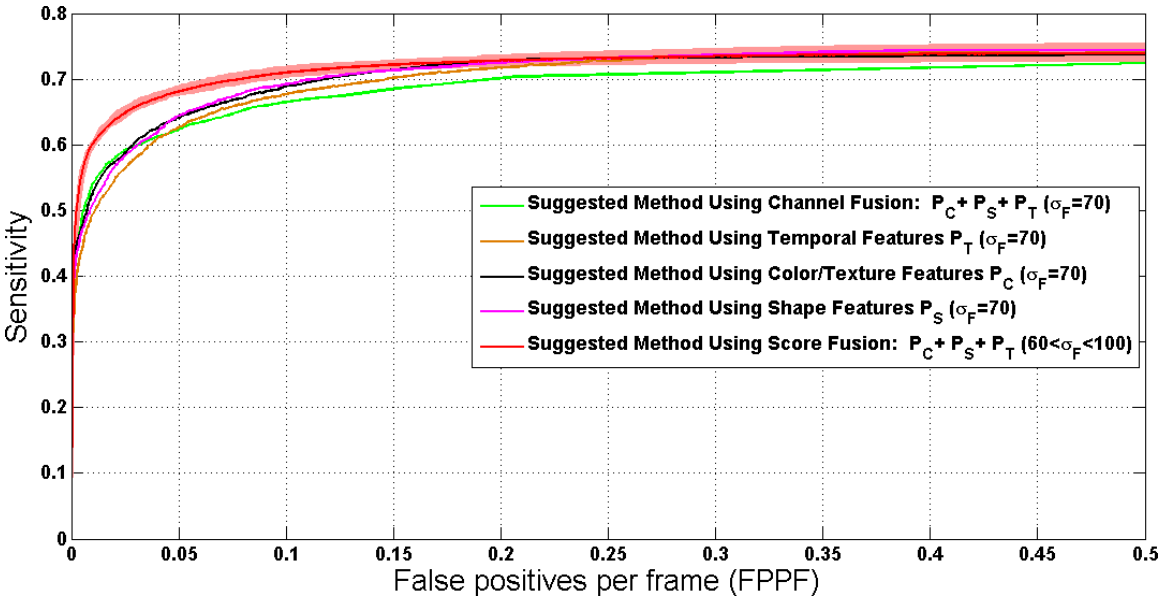


Figure 2.18: FROC analysis of our polyp detection system. The shaded FROC curve shows the variation in our system's performance when σ_F, which determines the size of voting fields, changes in a range of [60,100]. As seen, we observe a stable performance polyp detection performance over a wide range of voting fields. For comparison, we have included the best performance of the individual CNNs (obtained for σ_F=70) and their combination using a channel fusion approach.

positives generated in a dataset at different operating points. FROC analysis was originally designed for CT datasets and mammograms where the number of false positives were divided by the number of patients. With slight changes in definitions, FROC can be extended to colonoscopy videos. Specifically, we replace lesion localization rate with polyp detection rate and divide the number of false positives by the total number of frames in the videos. We show sensitivity to polyps on the vertical axis and the average number of false positive per frame on the horizontal axis.

In Figure 2.18, we show the FROC curves of our polyp detection system. The shaded FROC curve shows the variation in our system's performance when σ_F, which

determines the size of voting fields, changes in a range of [60,100]. As seen, we observe a stable performance over a wide range of voting fields. Recall that, in Section 2.4.1, we were not able to analyze the effect of σ_F on system's performance using a shape generator model; however, this analysis clearly shows the robustness of our system with respect to the changes in the size of voting fields. Our suggested system generates 0.002 false positives per frame at 50% sensitivity, which outperforms the system designed by Wang *et al.* (2013) that produces 0.15 false positives per frame at the same sensitivity.

To demonstrate the effectiveness of the suggested score fusion scheme (see Figure. 2.14), we have included FROCs for the individual CNNs trained on color patches (P_C), temporal patches (P_T), and shape in context patches (P_S), and their combination using a channel fusion approach. Channel fusion is performed by stacking color, shape, and temporal patches for each polyp candidate followed by training one CNN. This is in contrast to the score fusion approach (our main proposal), which as shown in Figure. 2.14, that involves three distinct CNNs each specialized in one type of features. As seen in Figure 2.18, the proposed score fusion framework significantly outperforms the performance of each individual CNN as well as their combination using the channel fusion approach.

Detection latency analysis. A major shortcoming of the FROC analysis is that it excludes the factor of time. While it is desirable for a computer-aided polyp detection system to localize as many instances of a polyp as possible, it is also important to measure how quickly the first instance of the polyp is detected by the CAD system, because the longer the polyps stay in the colonoscopic view, the more likely the colonoscopists can detect them on their own. We therefore propose a new performance curve that is more amenable to colonoscopy. Specifically, we measure the detection latency $\Delta T = \frac{t_2 - t_1}{fps}$ with respect to the number of false positives per frame, where

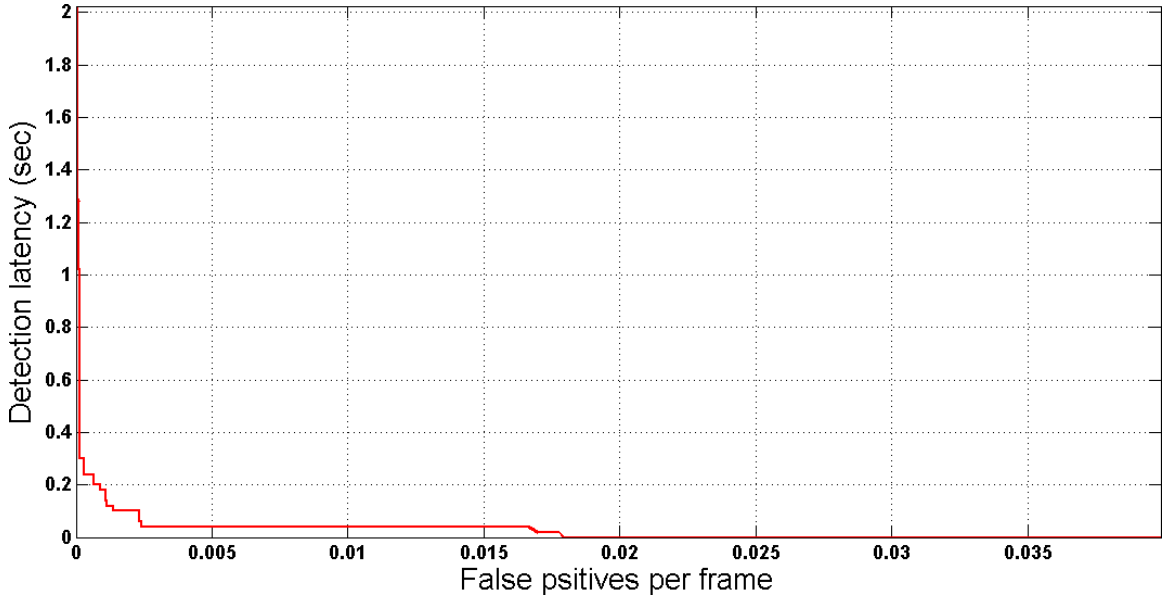


Figure 2.19: Detection latency analysis of our polyp detection system.

t_1 is the arrival frame of the polyp, t_2 is the frame in which the polyp is detected, and fps is the frame rate of the video. As with FROC, we change a threshold on the detection results and then at each operating point measure the median polyp detection latency of the test positive shots and the number of false positives in the entire test set. In Figure 2.19, we show the latency analysis of our system. As seen, our system can detect polyps in a fraction of second after they appear in the videos while generating a very few number of false positives.

2.4.3 Performance Comparison

In Table 2.3, we have summarized the recent polyp detection methods for colonoscopy videos. As seen, the existing systems have used different performance metrics for evaluation; some works use ROC plots, some employ FROC plots, and still others use the number of false positives per shot. Also as seen in Table 2.3, except for Tajbakhsh *et al.* (2013) and Bernal *et al.* (2012) that use a common public database *CVC-*

ColonDB, the rest perform evaluations using their own private databases. Given these limitations, we could make fair comparisons against only the works evaluated based on the public database. The competency of our approach against the other systems can be evaluated merely based on the reported results considering that they are obtained using different datasets. At the time of writing, the most recent CAD system for colonoscopy Wang *et al.* (2013) achieves a sensitivity of 34% at 0.1 false positive per frame, which is significantly outperformed by the suggested system with a sensitivity of 71% at the same false positive rate.

In Table 2.4, we compare the precision and recall rates of the proposed method with that of Bernal *et al.* (2012) and our early work using *CVC-ColonDB* . As seen, our method outperforms the method suggested by Bernal *et al.* (2012) with a large margin and also improves the precision of our previously published work. It is important to note that our polyp detection criterion is more strict than the one used by Bernal *et al.* (2012). We consider a polyp as "detected" if the polyp candidate falls inside the ground truth; however, they use a region-based approach that does not require the polyp candidates to fall inside the provided truth.

Theoretically, our CAD system is also expected to outperform the work of Bernal *et al.* (2012). To clarify this, we train the following binary classifiers: vessel vs. polyps, lumen areas vs. polyps, and specular reflections vs. polyps. Like before, we use oriented patches such that the structure of interest appear on right side of the patches and the background on the left side of the patches. We then obtain the variable importance map for each classification scenario (see Figure 2.20). Bernal *et al.* (2012) use the valley information for polyp localization, which corresponds to the information in the middle of the oriented patches. However, as seen in Figure 2.20, features from other parts of the patches are equally or even more important for discriminating polyps against vessels, specular reflections, and lumen areas. Relying

Table 2.3: Performance comparison between our system and the recent polyp detection methods designed for optical colonoscopy. The tabulated results are collected from the corresponding manuscripts. As seen, the existing methods are evaluated using different datasets and their results are reported based on different performance metrics (FPPF: false positives per frame, FPR: false positive rate, FPPS: false positives per shot). Our work can be fairly compared against Tajbakhsh *et al.* (2013) and Bernal *et al.* (2012) because all the three methods have been evaluated using the same public dataset.

Author	**Feature**	**Dataset** # images	**Result**
Current work	A comprehensive set of features	300 18900	86% recall @ 0.1 FPPF 71% recall @ 0.1 FPPF
Wang *et al.* (2013)	Edge cross section profile	1513	34% recall @ 0.1 FPPF
Tajbakhsh *et al.* (2013)	Shape in context	300	72% recall @ 10% FPR
Bernal *et al.* (2012)	Valley information	300	50% recall @ 10% FPR
Park *et al.* (2012)	Temporal and appearance features	35 videos	56% recall @ 10% FPR
Cheng *et al.* (2011)	Texture features	37	86.2% recall @ 1.26 FPPF
Ameling *et al.* (2009)	Texture and color features	1736	56% recall @ 10% FPR
Hwang *et al.* (2007)	Temporal and elliptical shape features	1 video	96% recall @ 1 FPPS

on only the central features results in weaker discrimination. Their work further assumes polyps as circular structures whose radii vary in a pre-specified range, but we do not assume any prior shape models for polyps and estimate the scales of polyps automatically.

On a desktop computer with a 2.4 GHz Intel quad core processor and an Nvidia GeForce GTX 760 video card, our system processes each image at 2.65 seconds, which is significantly faster than the system designed by Wang *et al.* (2013) with run-time

Table 2.4: Performance comparison based on *CVC-ColonDB*. Our system excels in all the operating points.

Recall	**Precision**		
	Our method	Tajbakhsh *et al.* (2013)	Bernal *et al.* (2012)
50%	100%	90%	92%
60%	98%	88%	78%
70%	95%	89%	65%
80%	92%	86%	60%

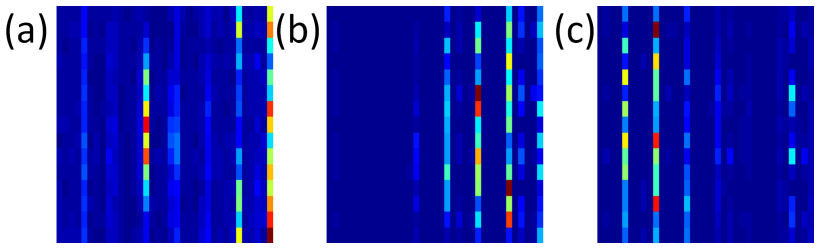


Figure 2.20: Variable importance maps generated by random forest classifiers for classifying polyps against (a) vessels, (b) lumen areas, (c) specular reflections. As seen, features from all over the patches are needed to distinguish polyps from the other structures in the images.

of 7.1 seconds and the system by Bernal *et al.* (2012) with run-time of 19 seconds. We should note that a very large fraction of the computation time (2.6 seconds) is caused by the candidate generation stage and that the candidate classification based on CNNs is extremely fast because CNNs are only applied to the candidate location in each frame. We expect a significant speedup of our system using parallel computing optimization.

2.5 Conclusion

In this chapter, we proposed a new computer-aided polyp detection system for colonoscopy videos. Our system began by generating a set of polyp candidates using context-aware shape features followed by an ensemble of convoluational neural networks to reduce the generated false positives. Our novel system consists of several new components: (1) a new patch descriptor, (2) a new edge classification scheme, (3) a new voting scheme, and (4) a new image presentation coupled with convolutional neural networks for false positive reduction. We evaluated our system using two polyp databases: (1) a public polyp database, *CVC-ColonDB*, containing 300 colonoscopy images with a total of 300 polyps, and (2) our collection of colonoscopy videos containing 18,900 frames and a total of 5,700 frames with polyps from 20 different exams. Our evaluations based on the public database demonstrated the superiority of the suggested system over the system designed by Bernal *et al.* (2012) where a significant improvement in precision was observed in nearly all operating points. To evaluate our system using our collection of colonoscopy videos, we employed the widely-used FROC analysis, achieving 50% sensitivity at 0.002 FPs/frame, outperforming the system designed by Wang *et al.* (2013) with 0.15 FPs/frame at the same sensitivity. We also evaluated our system using a latency analysis, demonstrating a very low polyp detection latency particularly in low false positive rates.

Chapter 3

AUTOMATIC VIDEO QUALITY ASSESSMENT

Automatic video quality assessment can warn colonoscopists against hasty and low quality colon visualization, which according to Arnold *et al.* (2009) account for 25% of a colonoscopy video. The existing methods in the literature for colonoscopy image quality assessment do not yield high sensitivity and specificity. This chapter presents a simple yet accurate method that can monitor the quality of colonoscopy videos more reliably.

In this chapter, we first summarize the existing methods for image quality assessment in colonoscopy, then present our suggested video quality assessment system, and then describe our image database collected at Mayo Clinic in Arizona. Finally, we present our experimental results.

3.1 Related Work

Image quality assessment is a well-studied field. The existing methods can be divided into (1) full reference techniques (e.g., Moulden *et al.* (1990)) that require the availability of the undistorted reference image, and (2) reduced reference and no reference techniques whose vast majority (e.g., Chen and Bovik (2011); Zhu and Milanfar (2009); Sazzad *et al.* (2008)) are distortion-specific. Considering that there is no unique reference image in colonoscopy, full reference methods are not amenable to colonoscopy. No reference methods are not suitable for colonoscopy either, because colonoscopy videos exhibit a wide range of degradation factors that are not commonly found in other applications. For instance, colonoscopy videos contain (1) motion blurred images that are captured during spasms of the colon or rapid movement

of the camera in the colon, (2) nearly flat images that are captured during wall contact, and (3) poorly visible images that are captured when the camera lens is occluded by fecal content. The limitations of the existing methods in dealing with such distortion factors have motivated research on developing colonoscopy-specific image quality assessment methods.

Filip *et al.* (2012) proposed a simple method based on image variance and contrast to detect blurry and out-of-focus images. Oh *et al.* (2007) suggested a more sophisticated method based on gray level co-occurrence matrix (GLCM) in the Fourier domain. The main idea was to identify frequency patterns associated with non-informative images. Arnold *et al.* (2009) computed the l^2-norm of 2D discrete wavelet transform (DWT) coupled with a Bayesian classification approach for image quality classification. Park *et al.* (2011) trained a hidden Markov model using entropy-related features. Our experiments, however, reveal that these methods fail to achieve high sensitivity and specificity for our diverse set of informative and non-informative images. This motivates our research to develop a more effective image quality assessment for colonoscopy.

3.2 Proposed Video Quality Assessment Method

We propose a system that can monitor the quality of colonoscopy and warn colonoscopists against hasty examinations or poor colon visualization. We base our method on the observation that a poor colon examination shot consists of a large number of non-informative frames including blurry, out-of-focus images, or those captured during wall contact (see Figure 3.1). Our method assigns each colonoscopy image a quality or informativeness score. By monitoring the informativeness scores during a procedure, we can detect the onset of a hasty or low quality colon examination upon identifying a number of consecutive non-informative images. In addition to the warn-

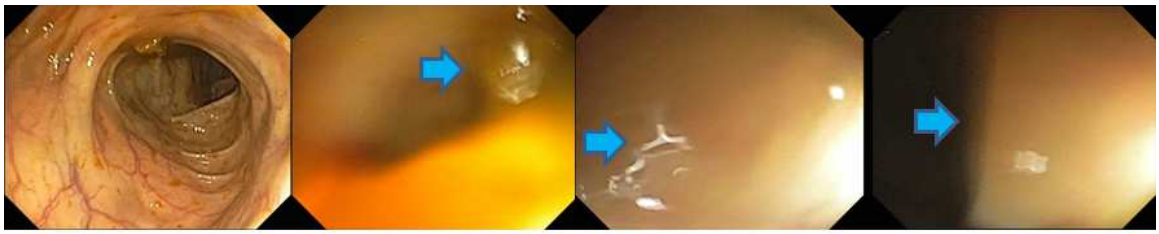


Figure 3.1: Comparison between informative and non-informative frames. (a) An informative frame where the information content is well-spread all over the image. (b-d) Examples of non-informative frames: (b) an out-of-focus image with bubbles inside, (c) an image captured during wall contact with light reflection artifacts, (d) a motion blurred image captured close to wall contact. Two distinguishing characteristics of non-informative frames are blurry edges and appearance of salient features in only a few local regions of the images (e.g., bubbles in (b) and reflection artifacts in (c)).

ing mechanism, the suggested method can report a segmental quality score during a procedure as the average of informativeness scores recorded in a segment of the colon. The overall quality score of a procedure can be computed as the average of the segmental scores, which can be used along with overall withdrawal time for quality monitoring purposes.

We base our methodology for detecting non-informative frames on two key observations (see Figure 3.1): (1) non-informative frames most often show an unrecognizable scene with few details and blurry edges and thus their information can be locally compressed in a few Discrete Cosine Transform (DCT) coefficients; however, informative images include much more details and their information content cannot be summarized by a small subset of DCT coefficients; (2) information content is spread all over the image in the case of informative frames, whereas in non-informative frames, depending on image artifacts and degradation factors, details may appear in

only a few regions. We use the former observation in designing our global features and the latter in designing our local image features.

Our method begins with dividing an input image to non-overlapping image patches. We then apply 2D DCT transform to each patch and reconstruct the patch using the dominant DCT coefficients. The reconstructed patches are then put together to form the reconstructed image. We create a difference map by taking the absolute difference between the original input image and the reconstructed image. The difference maps are computed in multiple scales, their histograms are constructed, and then concatenated to produce a global feature vector. We also compute a local feature vector to measure how information is spread all over the input image. To do so, we divide each difference map into a 3x3 grid and then in each cell, we compute the energy as the sum of squared intensities. The local feature vector is formed by concatenating all the local energy values computed in multiple scales. We experimentally found out that feature fusion is the best way to combine local and global information, outperforming other alternatives such as score-level and decision-level fusion. Once fused feature vectors are formed, we train a classifier to assign a probabilistic score to each input image with 0 and 1 indicating images with minimal and maximal information content, respectively. We refer to this score as "quality score" or "informativeness score" throughout the paper. We use random forest for classification. Random forest has been successfully applied to a variety of computer vision and medical image analysis applications, and has been shown to outperform other widely-used classifiers such as AdaBoost and support vector machines Criminisi and Shotton (2013). The two main ingredients of random forest are bagging of a large number of fully grown decision trees and random feature selection at each node while training the trees, which together achieve low generalization error and high quality probabilistic outputs. Figure 3.2 illustrates how the suggested method works. We should note that

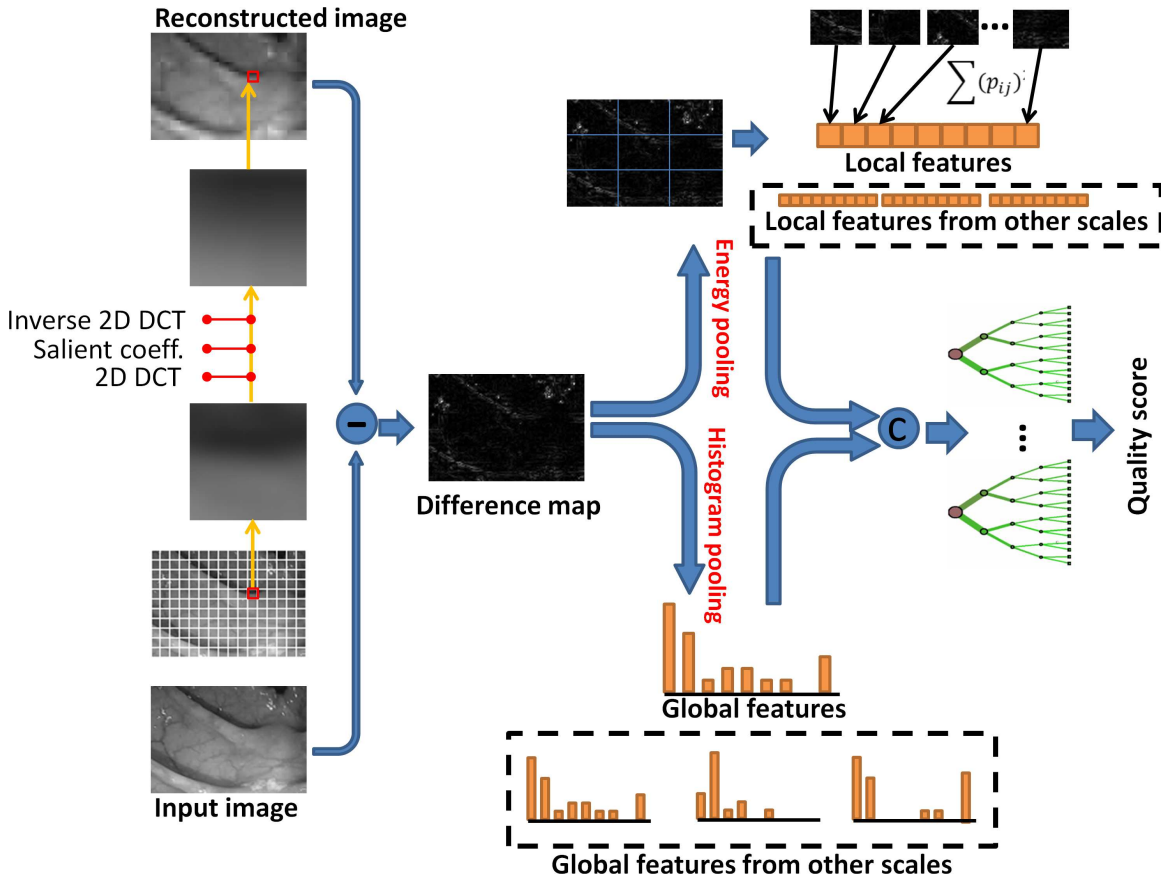


Figure 3.2: Overview of the suggested method for image informativeness assessment. Our method is based on global and local image features that are extracted by histogram pooling over the entire image reconstruction error and region-based energy pooling, i.e. l^2-norm of reconstruction error in each of the 9 sub-regions.

our method is fundamentally different from JPEG image compression. Unlike JPEG, our method is not meant to compress images, rather, it is designed to generate global and local feature vectors from the image reconstruction error.

3.3 Data

To evaluate our proposed system, we used six entire-length colonoscopy videos collected at Mayo Clinic in Arizona. Considering the expensive annotation of all

the video frames, we sampled each colonoscopy video by selecting one frame from every five seconds of the videos. By doing so, we removed many similar colonoscopy frames. The resulting set was further refined to make a balanced dataset where both the informative and non-informative classes were equally present. Our database of labeled colonoscopy images contain 5,500 colonoscopy images. A trained expert then manually labeled each frame in the database as informative or non-informative. The labeling of ambiguous frames was done after consulting with an experienced gastroenterologist. A limitation is that we did not have at least two experts to calculate inter-observer variability.

3.4 Performance Evaluation and Comparison

We divided the balanced dataset of 5,500 colonoscopy frames into a training set of 3,000 frames, a validation set of 1,000 frames, and a test set of 1500 frames. We used the validation set to tune the parameters of the method. We trained a random forest classifier consisting of 100 fully grown decision trees. We introduced randomness by randomly selecting a subset of features at the tree nodes.

We used receiver operating characteristic (ROC) curves to compare the performance of the suggested method with that of Oh *et al.* (2007) and Arnold *et al.* (2009). In Figure 3.3(a), we show the resulting ROC curves. As seen, our system outperforms the other two methods in all the operating points, achieving higher sensitivity and specificity to non-informative images. To test the statistical significance of the difference between the ROC curves, we employ the method designed by DeLong *et al.* (1988). Our statistical analysis shows that the proposed method achieves area under curve (AUC) of 0.948 (95% CI, 0.935 to 0.959), which significantly ($p < 0.0001$) outperforms Oh *et al.* (2007) with AUC of 0.880 (95% CI, 0.862 to 0.897) and Arnold *et al.* (2009) with AUC of 0.867 (95% CI, 0.848 to 0.885). For statistical analyses, we

used MedCalc for Windows, version 13.3 (MedCalc Software, Ostend, Belgium).

In Figure 3.3(b), we demonstrate image informativeness assessment for a segment of a colonoscopy video. Segments of the signal, highlighted with the ellipses, correspond to low quality colon examination because of their poor average quality score. For qualitative comparison, we selected three frames from this video and compared the corresponding informativeness scores assigned by our suggested system and the other methods. In Figure 3.3(c), we show a very informative colonoscopy frame, which is assigned the high score of 97% by our method, 62% by Oh *et al.* (2007), 89% by Arnold *et al.* (2009). In Figure 3.3(d), we show a rather blurry colonoscopy frame and the corresponding scores given by the three methods. As seen, the score assigned by our method (49%) describes the image quality more accurately. Finally, in Figure 3.3(e), we depict a non-informative colonoscopy frame captured during wall contact, to which our method assigns a poor score of 1% where as DWT gives a relatively high score of 53%. In summary, the scores generated by our system are more in-line with that of human perception.

On a desktop computer with a 2.4 GHz Intel quad core , our MATLAB implementation runs at 10 frame/sec, which outspeeds Oh *et al.* (2007) performing at 6.5 frame/sec. Compared with Arnold *et al.* (2009) that operates at 65 frame/second, our method is significantly slower; however, for real-time clinical applications, our method requires a speed-up of only 2.5, which is achievable using C++ Multi-threaded Programming.

3.5 Visual Feedback

Our method provides the operators with two types of visual feedback: (1) instant quality score, (2) average quality score. The instant quality score is shown as a bar with varying height in the top-left corner of the frames. A tall bar corresponds to an

informative frame with high quality colon visualization whereas a short bar indicates a non-informative frame with poor colon visualization. The average quality score is computed by averaging instant quality scores over every three seconds of the video and is shown in the form of a traffic light in the to-right corner of the frames. If the average score is high, the green light will be lit, indicating an examination with a high level of diligence. If the average score is medium, the yellow light will be lit, indicating an examination with a medium level of diligence. If the average score is low, the red light will be lit. A red light that continues for a few seconds triggers a warning message at the top of the screen, indicating an examination with poor level of diligence. In Figure 3.4, we provide examples of the visual feedback.

3.6 Conclusion

In this chapter, we presented a method for objective assessment of information content in colonoscopy images. The suggested method was designed according to two observations: (1) non-informative frames most often contain blurry edges; (2) information content is spread all over an image in the informative frames, whereas in non-informative frames, depending on image artifacts and degradation factors, details may appear in only a few regions. The former was modeled by pooling global features and the latter was captured by pooling local features from the absolute image reconstruction error. We suggested two visual feedback to summarize the results of objective quality assessment for colonoscopists. The instant quality was represented as a bar with varying height and the average quality of the video over the past three seconds as a traffic light. We trained and evaluated our method using our collection of 5,500 colonoscopy images. Our experiments demonstrated the superiority of the suggested method over the existing works both quantitatively and qualitatively.

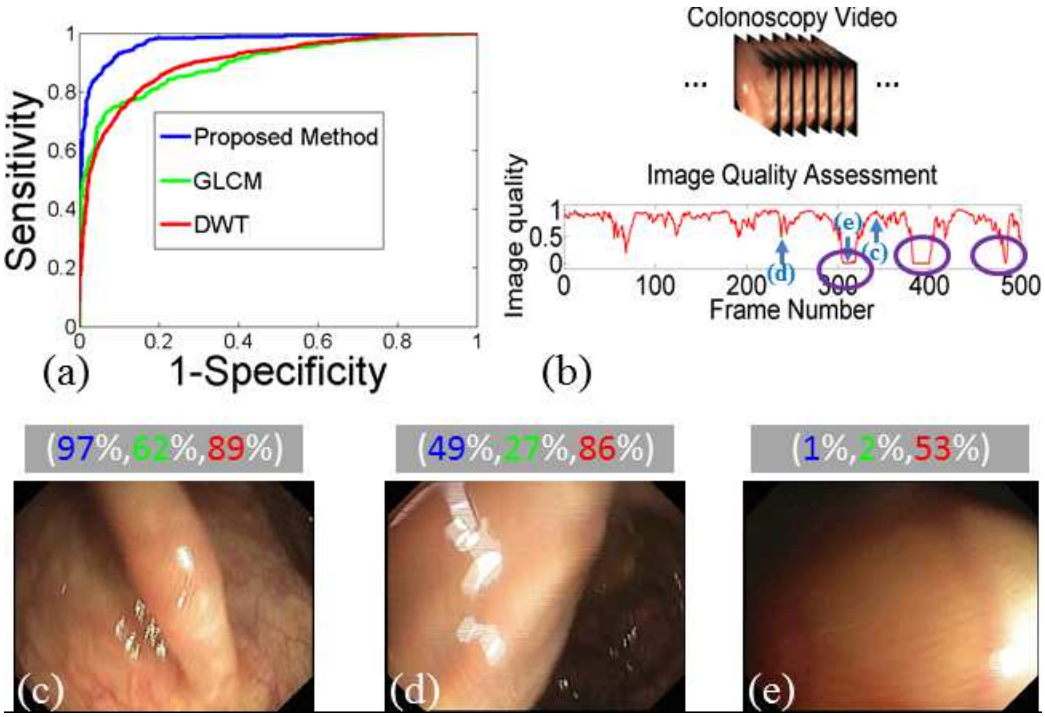


Figure 3.3: Quantitative and qualitative performance comparison. Comparison(a) Quantitative Comparison between the proposed method, DWT Arnold *et al.* (2009), and GLCM Oh *et al.* (2007). Our method excels in all operating points with a large margin. (b) Image informativeness assessment for a short colonoscopy video. The higher the score, the more informative the frame. Segments of the signal with average low quality scores correspond to hasty or low quality colon examination, in which case our method warns colonoscopists, encouraging a more diligent examination. (c-e) Three colonoscopy frames and their corresponding quality scores (our method in blue, GLCM in green and DWT in red). As seen, the scores assigned by our method are in more agreement with human perception.

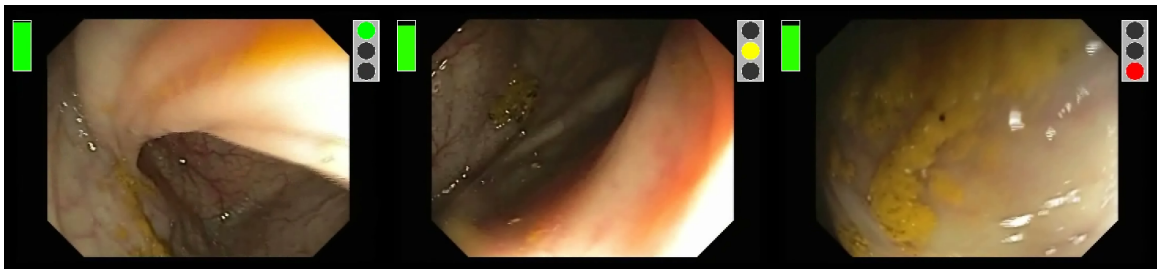


Figure 3.4: Examples of visual feedback generated by our quality assessment system. Three colonoscopy images with high instant quality (indicated by the green bar reaching its maximum height). Despite high quality, these three frames differ in the examination shot in which they were taken. (a) The green light indicates that the informative image is taken from a high quality examination shot. (b) The yellow light indicates that the informative image is taken from a mid-quality examination shot. (c) The red light indicates that the informative image is taken shortly after a poor examination shot because the red light is still on (average quality score is still low).

Chapter 4

3D VISUALIZATION OF THE COLON

A 3D visualization system can enhance the quality of colonoscopy by highlighting the regions that are missed during a procedure. Given a feedback on the location of the missed region, a colonoscopist can go back and search the area for polyps. In this chapter, we present a proof of concept of a 3D visualization system based on a depth camera and a tracking sensor.

This chapter begins with a review of the related work for 3D visualization of the colon. It then proceeds with explaining our suggested method followed by a proof of concept experiment. This chapter concludes with discussing the limitations of the suggested method for colonoscopy.

4.1 Related Work

3D visualization of gastrointestinal tract, and in particular the colon, is a very challenging task, which has been studied by only a few research teams. Deguchi (1996) proposed a 3D colon surface reconstruction technique using shape-from-shading (SfS). However, this work was mainly focused on computing 3D shapes of lesions such as polyps rather than building a 3D model of the entire colon. Koppel *et al.* (2007) employed Structure-from-Motion (SfM) to reconstruct static anatomic structures within the colon. Their method begins with extracting key points from colonoscopy frames. It then proceeds with establishing correspondences between the frames by matching the identified key points. The established correspondence are then used to project 2D information in the frames to the 3D space for surface reconstruction. However, as with Deguchi (1996), this work did not consider the 3D reconstruction of the colon rather

3D modeling of static anatomic structures. Kaufman and Wang (2008) proposed a reconstruction pipeline based on combined SfS and SfM [11]. Specifically, the authors suggested a 2-stage reconstruction method where the first stage built a partial colon surface using SfS but the second stage completed this model using information from SfM. However, this method was hindered by inherent limitation of SfM and SfS, that is inaccurate key point matching due to specular spots and non-rigid deformation of the colon.

3D colon reconstruction has recently been revisited by Hong *et al.* (2014). Considering the limitations of the previous approaches, they proposed a new reconstruction pipeline based on the colonic folds. Their method begins with identifying folds in colonoscopy images. For each fold, they estimate the distance to the camera, height of the fold, and width of the fold. The method then proceeds with projecting the folds into a 3D space. The projected folds serve as the skeleton of the colon, which can be used for 3D colon reconstruction. They qualitatively evaluated their system based on 12 colonoscopy frames taken from a phantom model of the colon. However, the success of this approach highly depends on an accurate identification of folds in the input images because mis-localization of the folds can distort the resulting 3D model of the colon.

In summary, despite different methodologies, the previous works have aimed at solving two common problems: (1) estimating the distance of the colonic structures with respect to the camera, (2) estimating the locations and orientations of the camera in the colon. For instance, Deguchi (1996) used SfS to estimate the depth or distance with respect to the camera. Koppel *et al.* (2007) and Kaufman and Wang (2008) employed SfM for localizing the camera in the colon and producing depth measurements. Hong *et al.* (2014) used changes in intensity values to estimate the depth of the colonic fold as well as their other physical measurements. Our research addresses

the problem of 3D visualization by exploring other alternatives to estimating depth and position information.

4.2 Proposed 3D Visualization System

We investigate a 3D visualization system that directly measures depth and position information using an additional set of hardware. This in contrast with the previous works where complicated and computationally expensive algorithms are used for estimating depth and position information. Thus, while the previous methods take a software-based approach for 3D reconstruction of the colon, we investigate the possibility of using a hardware-software approach for 3D colon reconstruction. Our hardware devices consists of (1) a tracking sensor that measures the location and orientation of the camera in real time, and (2) a depth camera that produces the depth images at 50 frames/second. Our software is a 3D reconstruction algorithm that fuses information from the tracking sensor and depth camera to build a 3D model of the environment. In the following, we first explain the employed hardware and then describe the 3D reconstruction algorithm.

4.2.1 Tracking Sensor

To obtain position and orientation of the camera, we use a tracking system manufactured by Ascension Technology Corporation [1] . Unlike the similar products that operate based on an AC magnetic field, Ascension Technology's tracking system uses a DC magnetic field, which offers new and improved ways to track six degrees-of-freedom sensors for high accuracy tracking of sensors, as small as a pencil point, in difficult metallic environments. This tracking system is fast, allowing for dynamic tracking with 240 to 420 measurements per second. In Figure 4.1, we show the track-

[1] http://www.ascension-tech.com/

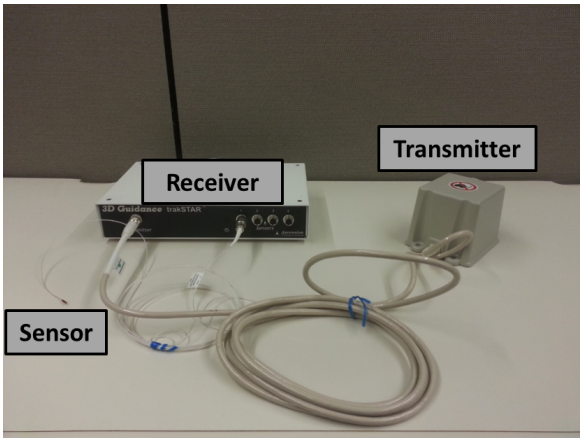


Figure 4.1: The tracking system we used in this study.

ing system that we have used in this study.

Our tracking system consists of three main components: (1) a transmitter unit that emits a weak magnetic field, (2) a small sensor unit whose response to the field varies as it is moved through it (with six degrees-of-freedom), (3) a receiver unit which receives the tracking information from the sensor and then sends the de-noised information to a desktop computer through a USB port. In this study, we have used a mid-range transmitter that allows for accurate measurements in a radius of 58 cm. We have also used a Model 180 Sensor that has an outer diameter of 2mm. The length of the sensor is 9.9mm, which is placed on a capable of length 3.3m.

4.2.2 Depth Camera

To acquire depth information, we use Creative Senz3D^{TM} depth and gesture camera. We chose this camera because it was the most accurate camera for short range measurements at the time we initiated the project. This camera is designed to detect hand gestures and head movement of the user within 20 to 50cm from the camera, enabling interactive games and human-computer interactions. The camera is shipped with a software package that tracks hand and face movements based on the acquired

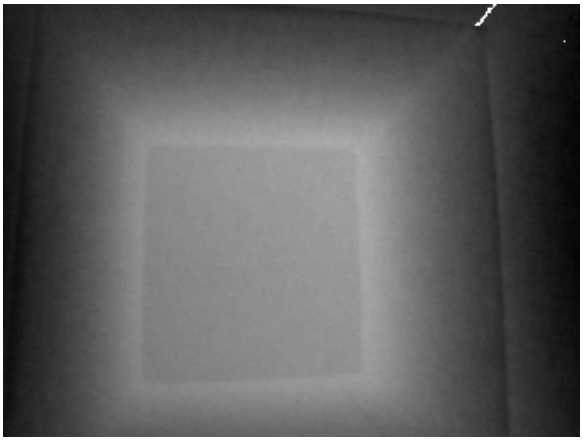


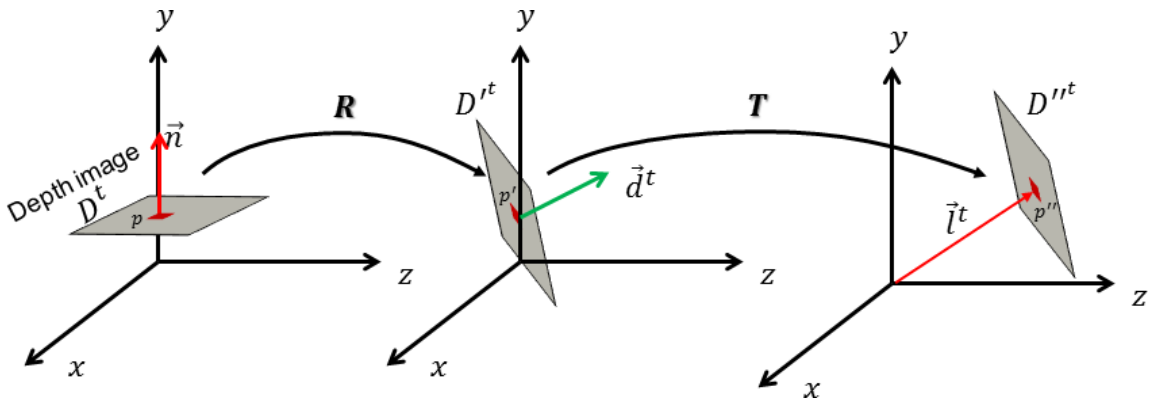
Figure 4.2: A depth image acquired by Creative Senz3D^{TM}. In a depth image, the objects that are closer to the camera appear darker than objects that are farther away from the camera.

depth images. In this project, we discard the software's tracking information and instead use the raw depth images.

Creative Senz3D^{TM} uses a time-of-flight (TOF) sensor, which illuminates the scene with a modulated light source, and observes the reflected light. The phase shift between the illumination and the reflection is measured and translated to distance for every pixel in a 2D addressable array, resulting in a depth map. Creative Senz3D^{TM} produces a 320 x 240 depth map at the rate of 50 frames/second. A depth map is essentially a gray scale image where a darker pixel intensity shows a relatively closer object to the camera while the brighter pixels indicate objects farther located from the camera. In Figure 4.2, we show a depth image captured by the depth camera.

4.2.3 3D Reconstruction Pipeline

In our 3D reconstruction pipeline, we frequently project a depth image to a cloud of points in 3D followed by a cloud matching process. We first provide a detailed explanation of these fundamental operations and then present the 3D reconstruction



(a)

Figure 4.3: Moving the depth image from a reference location to the target location by rotation R followed by translation T.

algorithm.

Projecting a depth image to a cloud of points in 3D. Here, we explain how to project a depth image captured at time t to a cloud of points in the 3D space. This process consists of 3 stages: (1) rotating the depth image around a rotation axis so that normal of the image plane becomes parallel to the camera axis, (2) translating the rotated depth image from a reference location to the target location according to the camera location, (3) projecting each pixel in the depth image to a point in a 3D point cloud.

Let D^t denote the depth image taken by the depth camera at time t. Also, let $\vec{d}^t$ and $\vec{l}^t$, respectively, denote the camera axis direction and camera location at time t. Without loss of generality, we assume that the depth image D^t is initially placed in the geometry shown in Figure 4.3, where the normal of the depth image $\vec{n}$ is parallel to the y-axis. In the first stage, we seek a rotation matrix R that can rotate $\vec{n}$ so that it becomes parallel to the camera axis $\vec{d}^t$. For this purpose, we first compute the rotation axis $\vec{r}$ as the cross product of the camera axis and the standard normal

direction:

$$\vec{r} = \vec{n} \times \vec{d^t} \tag{4.1}$$

The rotation matrix R for rotating a given vector around the rotation axis $\vec{r}$ in the amount of θ degree is then computed as follows:

$$R_{\vec{r}}(\theta) =$$

$$\begin{bmatrix} \cos\theta + r_x^2(1-\cos\theta) & -r_z\sin\theta + r_x r_y(1-\cos\theta) & r_y\sin\theta + r_x r_z(1-\cos\theta) \\ -r_z\sin\theta + r_x r_y(1-\cos\theta) & \cos\theta + r_y^2(1-\cos\theta) & r_x\sin\theta + r_y r_z(1-\cos\theta) \\ r_y\sin\theta + r_x r_z(1-\cos\theta) & r_x\sin\theta + r_y r_z(1-\cos\theta) & \cos\theta + r_z^2(1-\cos\theta) \end{bmatrix} \tag{4.2}$$

where

$$\sin\theta = \|\vec{v}_{ref} \times \vec{v}_{cam}\|, \cos\theta = \|\vec{v}_{ref} \cdot \vec{v}_{cam}\|$$

Once the rotation matrix is determined, we can rotate the depth image around the rotation axis $\vec{r}$ so that the normal of the depth image becomes parallel to the camera axis. Let p denote a pixel on the depth image D^t in the world coordinates where p is represented as a point in 3D space (the depth image is a plane in the 3D space). Our goal is to map this pixel to the desired location p' on the rotated depth image $D^{t\prime}$. If the world coordinates of p are known, then the solution will be as simple as applying a rotation in 3D space, $p' = Rp$ where R has been constructed above. For the geometry shown in Figure 4.3, the world coordinates of pixel p can be determined based on its image coordinates and the camera parameters as follows:

$$p = \begin{bmatrix} \tan(\alpha_h/2).f.(1-\frac{2x}{H}) \\ f \\ \tan(\alpha_w/2).f.(1-\frac{2z}{W}) \end{bmatrix} \tag{4.3}$$

Table 4.1: Parameters of the Senz3D camera retrieved from the API.

parameter	definition	value
α_h	angle of view along the height	58
α_w	angle of view along the width	74
H	height of the image in pixel	320
W	width of the image in pixel	240
f	focal length	22mm
Z_{min}	focal length	15cm
Z_{max}	focal length	100cm

where x and z denote the image coordinates of pixel p on the depth image D^t and the rest of the parameters are camera properties that are explained in Table 4.1. By applying the rotation matrix on all the pixels in the depth image D^t, we obtain the rotated depth image D'^t.

In the second stage, we seek a translation T that shifts the rotated depth image according to the camera location, resulting a new depth image D''^t. For this purpose, we translate all the pixel locations on the rotated depth image by $\vec{l^t}$, $p'' = p' + \vec{l^t}$ where $\vec{l^t}$ is the camera location at time t.

In the third stage, we perform projection to the cloud of points. To that end, for each pixel on D''^t, we obtain a unit vector whose extension connects the given pixel to the camera location. Such a unit vector for p'' can be computed as $\vec{u}_{p''} = \frac{p''-\vec{l^t}}{\|p''-\vec{l^t}\|}$. Next, pixel p'' gets projected to the cloud of points by a translation along the corresponding unit vector. Mathematically,

$$p''_C = p'' + d_{p''}\vec{u_{p''}} \tag{4.4}$$

where p''_C denotes the projected point in the cloud and $d_{p''}$ denotes the depth value

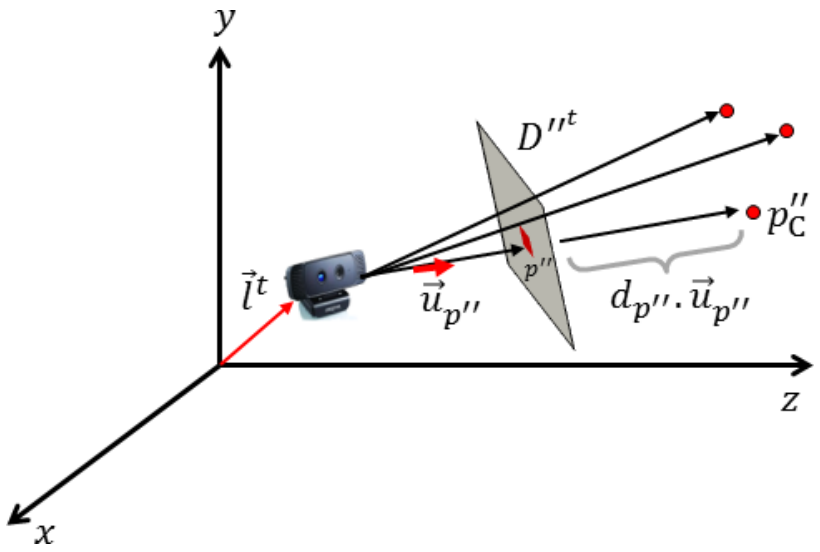


Figure 4.4: Projecting a depth image to a cloud of points in 3D.

for p''. The larger the depth value, the farther the point get projected. In Figure 4.4, we illustrate the projection process.

Cloud matching. Let P^t denote the cloud of points generated from depth image D^t that is taken at time t. The goal is to find a transformation T that aligns P^t with the existing cloud P. For this purpose, we employ point-to-plane cloud matching proposed by Chen and Medioni (1991).

The algorithm begins with estimating normal directions for each point p_i in the existing cloud P. Specifically, a number of nearby points are selected around p_i based on a kd-tree and then a principle component analysis is performed. The normal direction $\vec{n}_i$ at p_i is then selected as the eigen vector corresponding to the minimum resulting eigen value. Once normal directions are computed for all the points in P, we estimate the transformation T by minimizing the following constrained cost function:

$$\sum_{i=1}^{|P^t|}(Tq_i - p_j)\cdot\vec{n}_j, \text{with } p_j = \arg\max_{p\in P}\|Tq_i - p\| \tag{4.5}$$

where q_i is a point in P^t, p_j is the closest point in P with respect to q_i after

transformation T is applied, and $\vec{n}_j$ is the normal vector at p_j. As seen, the desired transformation T appears in both the cost function and the nonlinear constraint. One way to solve this optimization problem is to assume the transformation is known and then use the known transformation to compute the nearest neighbors p_js. Once the nearest neighbors are identified, one can cast away the constraint and re-compute T by minimizing the cost function using a least squares technique. The new transformation is then used to identify the nearest neighbors, which can be used to re-estimate the transformation matrix. This iterative process has proven to work robustly for cloud matching. The initial guess for T can be an identity transformation or the transformation that has been estimated for the depth image at time $t-1$.

3D reconstruction algorithm. The 3D reconstruction algorithm is split into two parts: (1) constructing a base cloud, (2) incrementally updating the base cloud. The base cloud is constructed based on the information acquired at time $t = 1$ and is then incrementally updated based on the information of time $t > 1$. In the following, we first explain how we construct the base cloud and then the mechanism we use for incremental updates. In Appendix C, we show the pseudocode of the suggested 3D reconstruction pipeline.

Let P denote the final cloud, which is an empty set of 3D points before the commencement of the 3D reconstruction. At time $t = 1$, we acquire the first depth image and the corresponding tracking information. We use these information to project the first depth image to a cloud of 3D points, P^1. Since the final cloud is currently empty, we add all the resulting 3D points in P^1 to the final cloud, $P = P \cup P^1$. We assign a variable importance, λ, to each point that is added to the final cloud, which indicates to what degree this point actually corresponds to a valid point in the 3D space. The variable importance is set to 1 for all the newly added points to the cloud.

At time $t > 1$, we acquire a depth image from the camera and position information from the tracking sensor. Following Eq. 4.4, we generate a point cloud P^t associated with the current depth image. However, it is always possible to have some degrees of cloud misalignment between the new cloud P^t and the existing cloud P. We therefore perform point cloud matching by finding a transformation T that maps P^t to a new point cloud P^t_T that is better aligned with the existing cloud (see Eq. 4.5). For cloud matching, we first find a subset of the points in the existing cloud P that is visible by the depth camera at time t. In Figure 4.5, we illustrate this by denoting in green the cloud seen by the camera. We then align the new point cloud with only the green subset of the existing cloud. By excluding the irrelevant (red) points, we can improve the accuracy of the cloud matching process.

Let p_i denote the i^{th} point in P^t_T and p_{nn} denote the nearest neighbor in the existing cloud to p_i. We compute the distance between these two points $d_i = \|p_{nn} - p_i\|$ and then perform one of the following actions:

1. if d_i is larger than a threshold (d_{MAX}), we conclude that the corresponding 3D point p_i is an outlier point, which may have been caused by inaccurate depth measurements. We therefore exclude p_i from the new point cloud P^t_T.

2. if d_i is smaller than a threshold (d_{MIN}), we conclude that p_i corresponds to a 3D point that already exist in the cloud P. We therefore remove p_i from the new cloud and increment the importance of p_{nn} by one unit. Note that existence of a point such as p_i very close to p_{nn} indicates that p_{nn} is indeed a valid point.

3. if d_i is between d_{MAX} and d_{MIN}, we conclude that p_i is neither an outlier point nor an existing point; rather, a new point that must be added to the existing cloud. We therefore keep p_i in P^t_T.

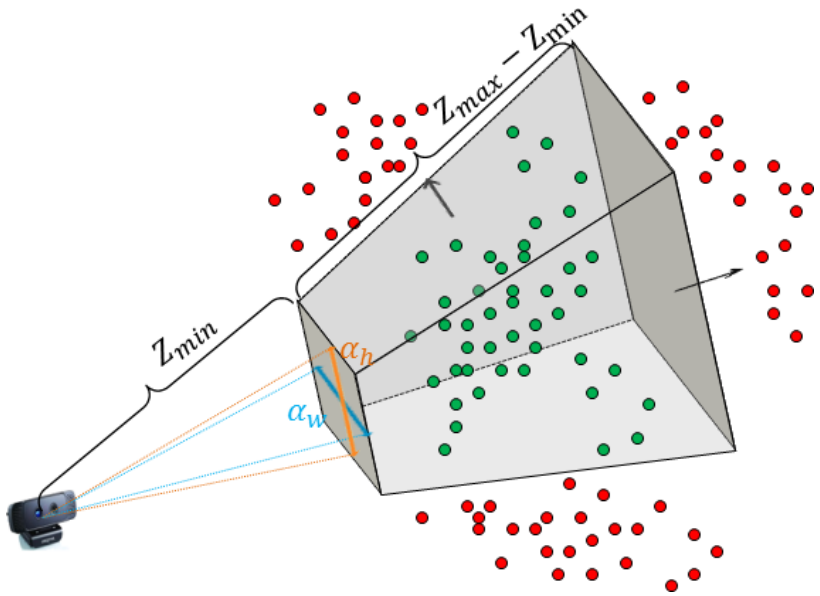


Figure 4.5: Cloud matching based on the camera information. The constructed cloud at time t can be divided into 2 subsets depending on the current location and orientation of the depth camera: (1) the green points that are visible to the camera, (2) and the red points that are not visible to the camera. Only green points are used for cloud registration. The pyramid that delineates the visibility range of the camera is computed based on the intrinsic parameters of the depth camera listed in Table 4.1 and the location and orientation of the camera given by the tracking sensor.

Once all the points in P_T^t are processed, we will add all the remaining points in P_T^t to the existing cloud. For model visualization, we plot only the 3D points of the existing cloud whose importance is larger than a previously set threshold.

4.2.4 A use case

As a proof of concept, we conduct the following experiment. We navigate the depth camera into a box so that the top side of the box always remain hidden to the camera. We then investigate whether the resulting 3D model can reveal that

the top side of the box has never been visited by the camera. To obtain positional information, we tape the tracking sensor on the depth camera such that the sensor orientation and camera axis become parallel. The navigation begins with placing the camera in a box followed by a gentle withdrawal. During withdrawal, we turn the camera to the left and right but ensure that the top side of the box always remains hidden to the camera. The resulting 3D cloud of points is shown in Figure 4.6. As seen, while the left, right and bottom side of the box is clearly constructed, the top side is missing, indicating that the camera has not visited the top portion of the box. This experiment demonstrates that the 3D reconstruction method coupled with a prior knowledge of the scene can reveal the region that has been missed during navigation.

4.2.5 Extension to Colonoscopy

Although the suggested method properly reconstructs a 3D model of the box, its extension to colonoscopy is hindered by a number of limitations. In the following, we outline these limitations and provide suggestions for future research.

A clean 3D reconstruction requires gentle movements of the camera in the colon. However, such a requirement is often not met in colonoscopy. This is because the human colon is a very uncontrolled environment, making it difficult to gently navigate the camera in the colon. Rapid movements of the camera will result in a distorted cloud of points, which will degrade the interpretability of the model for colonoscopists. The use of a tracking sensor can alleviate this problem to some degree but its full utility is limited by asynchronous acquisition rates of the tracking sensor and depth camera.

Effective feedback about the missed regions of the colon requires a previously-known geometry of the colon. Briefly, to identify the missed regions, one needs to

know both the geometry of the scene and the observed portion of the scene. Even if a reconstruction method succeeds in visualizing the visited part of the scene, it is nearly impossible to identify the unseen part with no prior geometric model of the scene. Note that even though the colon is generally a tubular structure, its geometric information such as orientation and diameter, and its intricacies such a height and locations of the folds are not the same throughout the entire colon.

The human colon is a non-rigid object, which makes both the 3D colon reconstruction and feedback process extremely difficult. The human colon constantly deforms due to spontaneous spasms of the colon. Such changes in the colon geometry can invalidate the reconstructed 3D model of the colon, weakening the correspondences between the model and the actual colon. Clearly, a system cannot provide meaningful feedback about the missed regions of the colon if the 3D model no longer matches the current geometry of the colon.

The size of the depth camera employed in this study hinders the evaluation of our method using a phantom model of the colon. The current depth cameras on the market are mainly designed for industrial robots, gaming applications, or facilitating human-computer interactions. For such applications, the depth cameras are usually manufactured in relatively large sizes; therefore, they cannot be employed for navigation in the colon or a phantom model of the colon. Another important consideration is that the operating range of the current depth cameras is between 20cm to a few meters, which is certainly not suitable for colonoscopy. Moreover, the current depth cameras are not suitable for reflective surfaces even if they appear in the operating range of the cameras. While this may not cause a problem when evaluation using a phantom model of the colon, extension to real colonoscopy is certainly problematic. This is because the colon surface is highly reflective, resulting in inaccurate depth maps and thus a distorted 3D model of the colon. However, most of these limitations

can be overcome by the advent of miniature stereo vision cameras. For instance, NanEye Stereo [2] is a stereo camera manufactured by AWAIBA Lda, which is of size 2.2x1x1.6mm and is based on a technology that is different from that of the existing depth cameras. The size and technology used in this camera make it amenable to colonoscopy.

4.3 Conclusion

In this chapter, we investigated a 3D reconstruction system based on a depth camera and a tracking sensor. Specifically, we used Creative Senz3D^{TM} camera for depth measurements and a tracking system manufactured by Ascension Technology Corporation for positional measurements. We presented a 3D reconstruction pipeline that utilized the information from the depth camera and tracking sensor for 3D model reconstruction. We explained each stage of the pipeline in details, providing both intuitive and mathematical explanation for each stage. In a use case experiment, we evaluated the suggested 3D reconstruction pipeline where the resulting model was able to reveal which part of the environment was hidden from the camera. However, the extension of our suggested 3D reconstruction pipeline to colonoscopy was challenging. The size and operating range of the existing depth cameras on one hand and the reflective surface of the colon on the other hand complicate the use of our system for 3D colon visualization. However, the recent advances in developing miniature stereo vision cameras can overcome these limitations, facilitating the extension of our 3D reconstruction system to colonoscopy.

[2] http://www.awaiba.com/product/naneye-stereo/

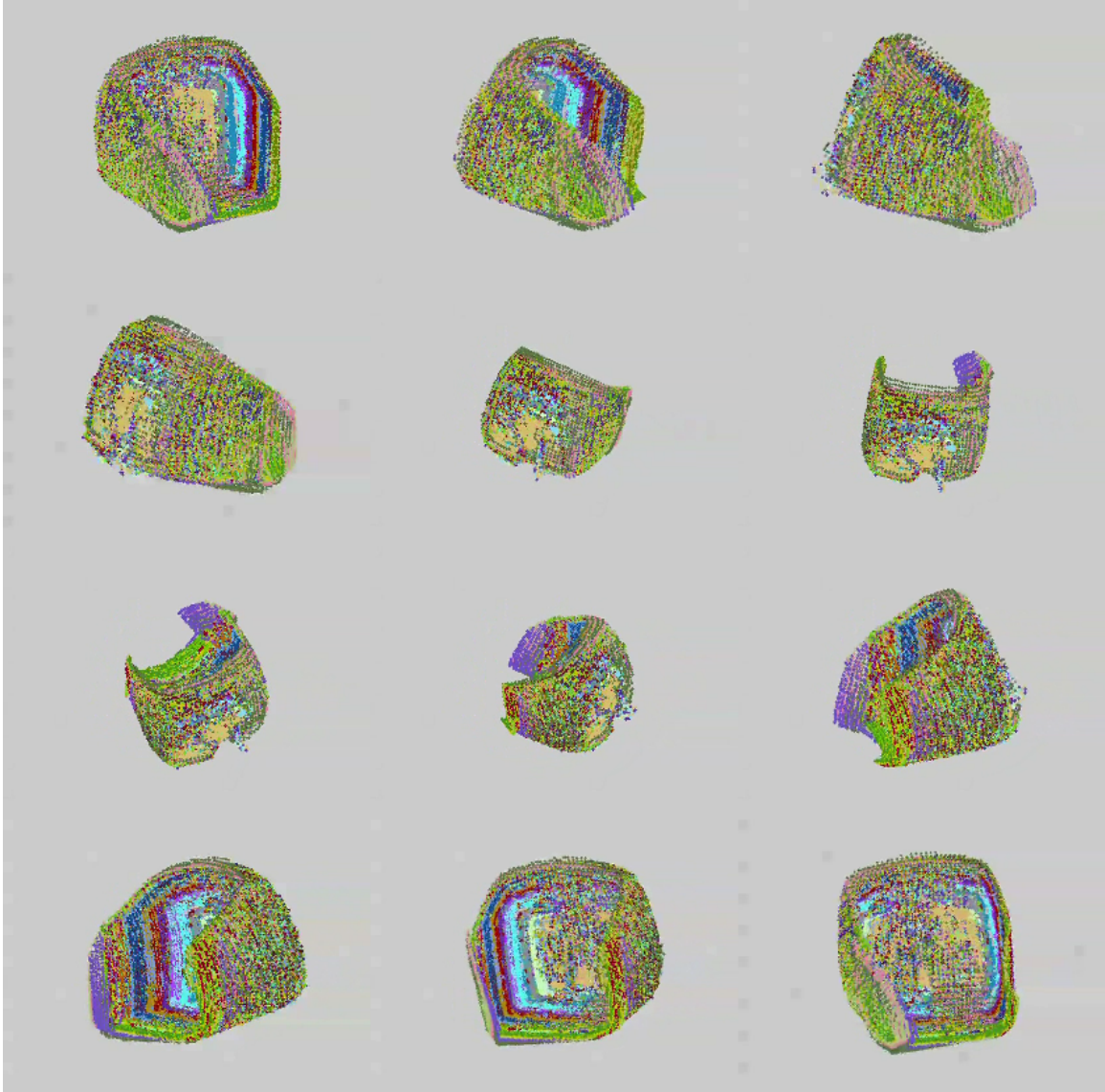


Figure 4.6: Different views of the 3D model constructed for a box. The camera is navigated in the box so that the top side of the box remains hidden to the camera. As seen, the model reveals that the top side of the box has not been captured by the camera. The color coding is used to show the contribution of each depth image to the final 3D model.

Chapter 5

DISCUSSION

We proposed a system with three components for ensuring high-quality colonoscopy: (1) an automatic polyp detection method, (2) an automatic quality assessment method, and (3) a 3D visualization method. In this chapter, we first explain the technical contributions of each method and then discuss the limitations and suggestions for future research.

5.1 Contributions

The suggested polyp detection method consists of several new components. **First**, a new patch descriptor that is designed to operate at high speed and low computational complexity. Our descriptor is both rotation invariant and robust against linear illumination changes. The former allows for capturing the patterns of intensity variation independent of boundary orientations. The latter upholds local invariance against varying lighting conditions. In addition, the suggested descriptor tolerates small degrees of positional changes, which is important to handle patch misalignment. **Second**, a two-stage classification framework that is able to enhance low level image features prior to classification. Unlike traditional image classification where a single patch undergoes the processing pipeline, our system fuses the information extracted from a pair of patches for more accurate edge classification. Together with the suggested patch descriptor, our classification scheme filters out the non-polyp edges from the edge maps. **Third**, a novel vote accumulation scheme that robustly detects polyps as objects with curvy boundaries in the refined edge maps. Our voting scheme does not require any predefined parametric model of shapes. As a byproduct, our vot-

ing scheme produces a bounding box around each polyp candidate, representing the size of a detected polyp. **Fourth**, a unique three-way image representation coupled with convolutional neural networks that allows us to learn a variety of polyp features such as color, texture, shape, and temporal information in multiple scales, enabling a more accurate polyp localization. Given a polyp candidate, a set of convolution neural networks—each specialized in one type of features—are applied in the vicinity of the candidate and then their results are aggregated to either accept or reject the candidate. The proposed system with the above components has yielded a superior performance over state of the art.

The suggested quality assessment system is based on a simple yet novel algorithm that can operate in real time. We based our algorithm on to two observations: (1) non-informative frames most often contain blurry edges; (2) information content is spread all over an image in the informative frames, whereas in non-informative frames, depending on image artifacts and degradation factors, details may appear in only a few regions. The former was modeled by pooling global features and the latter was captured by pooling local features from the absolute image reconstruction error. The local and global features are then used to build a classification model that assigns a normalized quality score to each colonoscopy frame. We also design two visual feedback for colonoscopists: (1) a bar with varying height to represent the instant image quality, and (2) a traffic light to represent the average quality scores and to warn against hasty or low quality colon visualization. Our experiments reveal that the suggested method achieves higher sensitivity and specificity to non-informative frames than the existing image quality assessment methods for colonoscopy videos.

The suggested 3D reconstruction system is based on a depth camera and a tracking sensor. We propose to use Creative Senz3D^{TM} camera for depth measurements and a tracking system manufactured by Ascension Technology Corporation for posi-

tional measurements. The employed hardware devices enable real time and accurate depth and position measurements. Our approach is fundamentally different from the similar works for colonoscopy where computationally expensive techniques such as structure from motion (SfM) are used to estimate depth and position information. Our 3D reconstruction pipeline is designed to utilize the information from the depth camera and tracking sensor for 3D model reconstruction. In a use case experiment, we evaluated the suggested 3D reconstruction pipeline and found out that the resulting model was able to reveal which part of the environment was hidden from the camera. We also discussed the limitations of the employed hardware for real colonoscopy and provided suggestions for future research.

5.2 Limitations and Future Work

Our polyp detection system can be improved in several ways. **First**, our system prototype currently processes each colonoscopy image in 2.65 seconds, which is not suitable for real time clinical use. While the current prototype can be significantly accelerated using C++ Multi-threaded Programming, further speedup can be achieved by changing the technology used in the suggested system. Specifically, future research can explore the possibility of replacing the candidate generation stage, which takes approximately 2.6 seconds, with a Hough transform. In doing so, one can obtain a relatively large number of polyp candidates in each colonoscopy image and then use convolutional neural networks for candidate classification. **Second**, our system is limited by detecting a maximum of one polyp in each colonoscopy image. At the first glance, it may not be problematic because a colonoscopy image is unlikely to contain more than a polyp; however, this can turn into a major limitation under special circumstances. Consider an image with a polyp in the center and a polyp-like structures in some other location in the image. Our system detects one polyp per image

and it is possible that this detection is cast in the region containing the polyp-like structure. Therefore, a polyp-like structure can mask the actual polyp. The future research can extend our voting scheme so that it can generate multiple detections per image. **Third**, our system has no mechanism to lock on the detected polyps. In other words, it is possible for our system to detect a polyp in frame t but fails to detect the same polyp in frame $t + 1$. The future work can research the possibility of simultaneous detection and tracking, which can significantly increase the sensitivity of polyp detection as well as the stability of the suggested system. **Fourth**, a thorough evaluation of the suggested system requires more data from different institutions. Our system is currently evaluated based on the largest available polyp database; however, a significantly larger database is needed to capture all variations in texture, shape, and color of polyps. In addition, all the videos used for performance evaluations have been collected in the department of Gastroenterology of Mayo Clinic in Arizona using Olympus colonoscopy machines. Such limitations in the number of videos and diversity of collecting institution can hinder the generalizability of our results.

Our quality assessment system yields high accuracy for detecting non-informative images; however, the evaluation phase can be significantly improved in terms of data, ground truth, and evaluation metrics. **First**, our current evaluations are based on 5,500 colonoscopy frames. The size of our database was limited by the manual and tedious process of labeling the colonoscopy images. Basically, each image in the database had to be labeled by an expert into informative and non-informative categories. While the labeling task was straightforward for a fraction of images, it was indeed challenging and time-consuming to label a large number of ambiguous colonoscopy images. **Second**, in the current study, all the images were labeled by a trained expert, which is certainly not the best practice given the subjectivity involved with labeling colonoscopy images. The key to high quality and noise-free image labels

is to label images based on consensus of a number of experts. **Third**, we currently evaluate our system based on how well it detects non-informative images. However, it will more desirable to investigate the correlation between the computed quality scores and the polyp detection rates or experience of colonoscopists. A significant correlation will indicate that our system can indeed evaluate the quality of colonoscopy procedures.

Our 3D visualization system performed desirably in a use case experiment; however, it can be significantly improved in terms of both algorithm and the hardware devices used in our system. **First**, our current method is based on a linear cloud matching algorithm, which is limited to matching clouds only if they are related through a linear transformation. Clearly, a cloud matching algorithm capable of finding nonlinear transformations can improve the quality of the resulting 3D model. A nonlinear cloud matching algorithm can alleviate the problem of reconstructing non-rigid objects such as the human colon. **Second**, the extension of the current 3D visualization system to colonoscopy is limited by the size, operating range, and technology of the depth camera that we used in this study. The existing depth cameras on the market are relatively large because they are primarily designed for gaming and other industrial applications. Cameras whose size is in the order of centimeters are undesirably large for colonoscopy. In addition, a depth camera for colonoscopy must provide accurate depth measurements for objects that appear within millimeters from the depth camera. However, the operating ranges of the existing short range depth cameras change between 20cm to several meters. Furthermore, a vast majority of the depth cameras operate based on the time of flight technology, which provides inaccurate depth measurements for reflective surfaces. Considering that the human colon is a highly reflective surface, the depth cameras based on the time of flight technology are not suitable for colonoscopy. However, most of these limitations can be overcome

by the recently released miniature stereo vision cameras, which are very small (in the order of millimeters) and operate based on a technology that is different from that of the depth cameras. The size and technology used in this camera make it amenable to colonoscopy.

REFERENCES

Alexandre, L. A., N. Nobre and J. Casteleiro, "Color and position versus texture features for endoscopic polyp detection", in "BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on", vol. 2, pp. 38–42 (IEEE, 2008).

Ameling, S., S. Wirth, D. Paulus, G. Lacey and F. Vilarino, "Texture-based polyp detection in colonoscopy", in "Bildverarbeitung fr die Medizin 2009", edited by H.-P. Meinzer, T. Deserno, H. Handels and T. Tolxdorff, Informatik aktuell, pp. 346–350 (Springer Berlin Heidelberg, 2009).

Arnold, M., A. Ghosh, G. Lacey, S. Patchett and H. Mulcahy, "Indistinct Frame Detection in Colonoscopy Videos", 2009 13th International Machine Vision and Image Processing Conference pp. 47–52 (2009).

Bernal, J., J. Snchez and F. Vilario, "Towards automatic polyp detection with a polyp appearance model", Pattern Recognition **45**, 9, 3166–3182 (2012).

Birkner, B., J. Knopnadel and L. A. et al., "Quality performance of screening colonoscopy in germany", Gastrointest Endosc. **61**, 5, p.AB248 (2005).

Bressler, B., L. F. Paszat, Z. Chen, D. M. Rothwell, C. Vinden and L. Rabeneck, "Rates of new or missed colorectal cancers after colonoscopy and their risk factors: A population-based analysis", Gastroenterology **132**, 1, 96 – 102 (2007).

Chen, M.-J. and A. C. Bovik, "No-reference image blur assessment using multiscale gradient", EURASIP Journal on image and video processing **2011**, 1, 1–11 (2011).

Chen, Y. and G. Medioni, "Object modeling by registration of multiple range images", in "Robotics and Automation, 1991. Proceedings., 1991 IEEE International Conference on", pp. 2724–2729 (IEEE, 1991).

Cheng, D.-C., W.-C. Ting, Y.-F. Chen and X. Jiang, "Automatic detection of colorectal polyps in static images", Biomedical Engineering: Applications, Basis and Communications **23**, 05, 357–367 (2011).

Ciresan, D., U. Meier and J. Schmidhuber, "Multi-column deep neural networks for image classification", in "Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on", pp. 3642–3649 (IEEE, 2012).

Criminisi, A. and J. Shotton, *Decision Forests for Computer Vision and Medical Image Analysis* (Springer Publishing Company, Incorporated, 2013).

Dalal, N. and B. Triggs, "Histograms of oriented gradients for human detection", in "Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on", vol. 1, pp. 886–893 vol. 1 (2005).

Deguchi, K., "Shape reconstruction from endoscope image by its shadings", in "Multisensor Fusion and Integration for Intelligent Systems, 1996. IEEE/SICE/RSJ International Conference on", pp. 321–328 (1996).

DeLong, E. R., D. M. DeLong and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach", Biometrics pp. 837–845 (1988).

Figueiredo, I. N., S. Kumar and P. N. Figueiredo, "An intelligent system for polyp detection in wireless capsule endoscopy images", Computational Vision and Medical Image Processing IV: VIPIMAGE 2013 p. 229 (2013).

Filip, D., X. Gao, L. Angulo-Rodríguez, M. P. Mintchev, S. M. Devlin, A. Rostom, W. Rosen and C. N. Andrews, "Colometer: a real-time quality feedback system for screening colonoscopy.", World journal of gastroenterology **18**, 32, 4270–7 (2012).

Frandzel, S., "P4p colonoscopy: An answer to falling medicare reimbursement rates?", Gastroenterology and Endoscopy News **58:09** (2007).

Gelder, R. E. V., C. Nio, J. Florie, J. F. Bartelsman, P. Snel, S. W. de Jager, S. J. van Deventer, J. S. Lamris, P. M. Bossuyt and J. Stoker, "Computed tomographic colonography compared with colonoscopy in patients at increased risk for colorectal cancer", Gastroenterology **127**, 1, 41–48 (2004).

Gross, S., S. Palm, J. J. W. Tischendorf, A. Behrens, C. Trautwein and T. Aach, "Automated classification of colon polyps in endoscopic image data", (2012).

Heresbach, D., T. Barrioz, D. Lapalus, MG.and Coumaros and et al., "Miss rate for colorectal neoplastic polyps: a prospective multicenter study of back-to-back video colonoscopies", Endoscopy **40**, 4, 284–290 (2008).

Hewett, D. G., C. J. Kahi and D. K. Rex, "Does colonoscopy work?", Journal of the National Comprehensive Cancer Network **8**, 1, 67–77 (2010).

Hong, D., W. Tavanapong, J. Wong, J. Oh and P. C. de Groen, "3d reconstruction of virtual colon structures from colonoscopy images", Computerized Medical Imaging and Graphics **38**, 1, 22–33 (2014).

Hwang, S. and M. Celebi, "Polyp detection in wireless capsule endoscopy videos based on image segmentation and geometric feature", in "Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on", pp. 678–681 (2010).

Hwang, S., J. Oh, W. Tavanapong, J. Wong and P. de Groen, "Polyp detection in colonoscopy video using elliptical shape feature", in "Image Processing, 2007. ICIP 2007. IEEE International Conference on", vol. 2, pp. II–465–II–468 (2007).

Iakovidis, D. K., 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 "Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on", pp. 575–580 (IEEE, 2005).

Karkanis, S. A., D. K. Iakovidis, D. E. Maroulis, D. A. Karras and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features", Information Technology in Biomedicine, IEEE Transactions on **7**, 3, 141–152 (2003).

Kaufman, A. and J. Wang, "3d surface reconstruction from endoscopic videos", in "Visualization in Medicine and Life Sciences", pp. 61–74 (Springer, 2008).

Koppel, D., C. Chen, Y. Wang, H. Lee, J. Gu, A. Poirson and R. Wolters, "Toward automated model building from video in computer-assisted diagnoses in colonoscopy", (SPIE - The International Society of Optical Engineering, 2007).

Krizhevsky, A., I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", in "Advances in neural information processing systems", pp. 1097–1105 (2012), available at `https://code.google.com/p/cuda-convnet/`.

Kwitt, R., N. Rasiwasia, N. Vasconcelos, A. Uhl, M. Hfner and F. Wrba, "Learning pit pattern concepts for gastroenterological training", in "Medical Image Computing and Computer-Assisted Intervention MICCAI 2011", edited by G. Fichtinger, A. Martel and T. Peters, vol. 6893 of *Lecture Notes in Computer Science*, pp. 280–287 (Springer Berlin Heidelberg, 2011).

Lebwohl, B., F. Kastrinos, M. Glick, A. J. Rosenbaum, T. Wang and A. I. Neugut, "The impact of suboptimal bowel preparation on adenoma miss rates and the factors associated with early repeat colonoscopy", Gastrointestinal endoscopy **73**, 6, 1207–1214 (2011).

Lee, J.-G., J. H. Kim, S. H. Kim, H. S. Park and B. I. Choi, "A straightforward approach to computer-aided polyp detection using a polyp-specific volumetric feature in {CT} colonography", Computers in Biology and Medicine **41**, 9, 790 – 801 (2011).

Lee, T. J. W., M. D. Rutter, R. G. Blanks, S. M. Moss, A. F. Goddard, A. Chilton, C. Nickerson, R. J. Q. McNally, J. Patnick and C. J. Rees, "Colonoscopy quality measures: experience from the nhs bowel cancer screening programme", Gut **61**, 7, 1050–1057 (2012).

Lieberman, D., M. Nadel, R. A. Smith, W. Atkin, S. B. Duggirala, R. Fletcher, S. N. Glick, C. D. Johnson, T. R. Levin, J. B. Pope, M. B. Potter, D. Ransohoff, D. Rex, R. Schoen, P. Schroy and S. Winawer, "Standardized colonoscopy reporting and data system: report of the quality assurance task group of the national colorectal cancer roundtable", Gastrointestinal Endoscopy **65**, 6, 757–766 (2007).

Mohamed, R., A. A. Shaheen and M. Raman, "Evaluation of colonoscopy skills–how well are we doing?", Canadian Journal of Gastroenterology **25**, 4, 198 (2011).

Mordohai, P. and G. Medioni, *Tensor Voting: A Perceptual Organization Approach to Computer Vision And Machine Learning*, Synthesis Lectures on Image, Video, And Multimedia Processing (Morgan & Claypool Publishers, 2007).

Moulden, B., L. F. Gatley *et al.*, "The standard deviation of luminance as a metric for contrast in random-dot images.", Perception (1990).

Oh, J., S. Hwang, J. Lee, W. Tavanapong, J. Wong and P. C. de Groen, "Informative frame classification for endoscopy video.", Medical image analysis **11**, 2, 110–27 (2007).

Ojala, T., M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns", Pattern Analysis and Machine Intelligence, IEEE Transactions on **24**, 7, 971–987 (2002).

Ong, J. and A.-K. Seghouane, "From point to local neighborhood: Polyp detection in ct colonography using geodesic ring neighborhoods", Image Processing, IEEE Transactions on **20**, 4, 1000–1010 (2011).

Park, S. Y., D. Sargent, I. Spofford and K. Vosburgh, "Colonoscopy video quality assessment using hidden markov random fields", in "SPIE Medical Imaging", pp. 79632P–79632P (International Society for Optics and Photonics, 2011).

Park, S. Y., D. Sargent, I. Spofford, K. Vosburgh and Y. A-Rahim, "A colon video analysis framework for polyp detection", Biomedical Engineering, IEEE Transactions on **59**, 5, 1408–1418 (2012).

Rabeneck, L., H. El-Serag, J. Davila and R. Sandler, "Outcomes of colorectal cancer in the united states: no change in survival (1986-1997).", The American journal of gastroenterology **98**, 2, 471 (2003).

Raman, M. and T. Donnon, "Procedural skills education–colonoscopy as a model", Canadian Journal of Gastroenterology **22**, 9, 767 (2008).

Rex, D., C. Cutler, G. Lemmel, E. Rahmani and et al., "Colonoscopic miss rates of adenomas determined by back-to-back colonoscopies", Gastroenterology **112**, 1, 24–28 (1997).

Rex, D. K., "Colonoscopic withdrawal technique is associated with adenoma miss rates", Gastrointestinal endoscopy **51**, 1, 33–36 (2000).

Rex, D. K., "Maximizing detection of adenomas and cancers during colonoscopy", Am J Gastroenterol **101**, 12, 2866–2877 (2006).

Rex, D. K., J. H. Bond, S. Winawer, T. R. Levin, R. W. Burt, D. A. Johnson, L. M. Kirk, S. Litlin, D. A. Lieberman, J. D. Waye, J. Church, J. B. Marshall and R. H. Riddell, "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", Am J Gastroenterol **97**, 6, 1296–308 (2002).

Robertson, D. J., "Colonoscopy for colorectal cancer prevention: is it fulfilling the promise?", Gastrointestinal Endoscopy **71**, 1, 118–120 (2010).

Robertson, D. J., D. A. Lieberman, S. J. Winawer, D. Ahnen, E. R. Greenberg, J. A. Baron, J. H. Bond, R. Jiang and M. E. Martinez, "795 interval cancer after total colonoscopy: Results from a pooled analysis of eight studies", Gastroenterology **134**, 4, Supplement 1, A–111 – A–112 (2008).

Saini, A., P. Bekal, A. Safdi, D. Walker, B. Saini, M. Chokshi and M. Safdi, "Effect of real-time visual feedback on adenoma detection rate", in "2009 ACG Abstracts", (2009).

Sazzad, Z. P., Y. Kawayoke and Y. Horita, "No reference image quality assessment for jpeg2000 based on spatial features", Signal Processing: Image Communication **23**, 4, 257–268 (2008).

Siegel, R. L., K. D. Miller and A. Jemal, "Cancer statistics, 2015", CA: a cancer journal for clinicians **65**, 1, 5–29 (2015).

Silva, J., A. Histace, O. Romain, X. Dray and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer", International Journal of Computer Assisted Radiology and Surgery pp. 1–11 (2013).

Suzuki, K., J. Zhang and J. Xu, "Massive-training artificial neural network coupled with laplacian-eigenfunction-based dimensionality reduction for computer-aided detection of polyps in ct colonography", Medical Imaging, IEEE Transactions on **29**, 11, 1907–1917 (2010).

Tajbakhsh, N., S. Gurudu and J. Liang, "A classification-enhanced vote accumulation scheme for detecting colonic polyps", in "Abdominal Imaging. Computation and Clinical Applications", vol. 8198 of *Lecture Notes in Computer Science*, pp. 53–62 (2013).

Tamaki, T., J. Yoshimuta, M. Kawakami, B. Raytchev, K. Kaneda, S. Yoshida, Y. Takemura, K. Onji, R. Miyaki and S. Tanaka, "Computer-aided colorectal tumor classification in {NBI} endoscopy using local features", Medical Image Analysis **17**, 1, 78 – 100 (2013).

Tola, E., V. Lepetit and P. Fua, "DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo", IEEE Transactions on Pattern Analysis and Machine Intelligence **32**, 5, 815–830 (2010).

Van Ravesteijn, V., C. van Wijk, F. Vos, R. Truyen, J. Peters, J. Stoker and L. van Vliet, "Computer-aided detection of polyps in ct colonography using logistic regression", Medical Imaging, IEEE Transactions on **29**, 1, 120–131 (2010).

van Rijn, J., J. Reitsma, J. Stoker, P. Bossuyt, S. van Deventer and E. Dekker, "Polyp miss rate determined by tandem colonoscopy: a systematic review", American Journal of Gastroenterology **101**, 2, 343–350 (2006).

van Wijk, C., V. Van Ravesteijn, F. Vos and L. van Vliet, "Detection and segmentation of colonic polyps on implicit isosurfaces by second principal curvature flow", Medical Imaging, IEEE Transactions on **29**, 3, 688–698 (2010).

Wang, Y., W. Tavanapong, J. Wong, J. Oh and P. de Groen, "Part-based multi-derivative edge cross-section profiles for polyp detection in colonoscopy", Biomedical and Health Informatics, IEEE Journal of **PP**, 99, 1–1 (2013).

Zhao, L., C. Botha, J. Bescos, R. Truyen, F. Vos and F. Post, "Lines of curvature for polyp detection in virtual colonoscopy", IEEE Transactions on Visualization and Computer Graphics **12**, 5, 885–892 (2006).

Zhu, H., Y. Fan and Z. Liang, "Improved curvature estimation for shape analysis in computer-aided detection of colonic polyps", in "Virtual Colonoscopy and Abdominal Imaging. Computational Challenges and Clinical Opportunities", vol. 6668, pp. 9–14 (Springer Berlin Heidelberg, 2011).

Zhu, X. and P. Milanfar, "A no-reference sharpness metric sensitive to blur and noise", in "Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on", pp. 64–69 (IEEE, 2009).

APPENDIX A

CANDIDATE GENERATION METHOD

The pseudocode of the suggested polyp candidate generation method.

Input:

- A colonoscopy image I
- Trained random forest classifiers $RF_i|_{i=1}^2$

Output:

- A polyp candidate

Candidate generation process

{**Step 1: Collect edges and normals**}
$E = \{(e_i, n_i) \ | \angle n_i \in [0, \pi), i = 1, 2, ..., N\}$

{**Step 2: Refine the edge map via classification**}
for $i = 1, 2, ..., N$ //for each edge

{**Step 2.1: Extract a pair of patches**}
$\{n_i^1 \leftarrow n_i, n_i^2 \leftarrow -n_i\}$ //assuming two normals
p_i^1 oriented patch with n_i^1 being the normal
p_i^2 oriented patch with n_i^2 being the normal

{**Step 2.2: Extract features** }
$d_i^1 \leftarrow \mathcal{F}(p_i^1)$, $d_i^2 \leftarrow \mathcal{F}(p_i^2)$

{**Step 2.3: Classify edges**}
{**A. Generate mid-level features** }
$f_i^1 \leftarrow RF_1(d_i^1)$, $f_i^2 \leftarrow RF_1(d_i^2)$
$f_i \leftarrow c(f_i^1, f_i^2)$ //concatenation
{**B. Fuse patch information**}
S$\leftarrow RF_2(f_i)$ //1x3 array
if $S[1] > 0.5$
$y_i \leftarrow 1$, $n_i^* \leftarrow n_i^1$ //edge accepted
else if $S[2] > 0.5$
$y_i \leftarrow 2$, $n_i^* \leftarrow n_i^2$ //edge accepted
else
$y_i \leftarrow 0$ //edge rejected
end if
end for

{**Step 3: Localize polyps through voting**}
{**Step 3.1: Group edges**}
$V^k = \{e_i | y_i \notin 0 \wedge \frac{k\pi}{4} < mod(\angle n_i^*, \pi) < \frac{(k+1)\pi}{4}\}$
{**Step 3.2: Candidate generation**}
$\mathcal{M}^k = \sum_{v_i \in V^k} M_{v_i}(x, y), k = 0, 1, 2, 3$

$candidate \leftarrow \underset{x,y}{\text{argmax}} \prod_{k=0}^{3} \mathcal{M}^k$

APPENDIX B

EDGE CLASSIFICATION SCHEME

This pseudocode explains how the suggested edge classification pipeline is trained.

Input:

- A set of training images $\mathcal{I} = \{I^1, I^2, ..., I^m\}$
- Ground truth images $\mathcal{G} = \{G^1, G^2, ..., G^m\}$
  - Truth make up $G(x, y) \in \{1, 2, 3, 4, 5\}$

1: polyp, 2: vessel, 3: lumen, 4: specular reflection, 5: random

Output:

- Trained random forest classifiers RF_1, RF_2

Learning edge classifiers

{**Layer 1: Train the 1st classifier**}

Step0: collect labeled edges

E={}//set of edges
L={}//set of labels
N={}//set of desired normals
for i=1...m //for each image
$I_{bin} = edge(I^i)$
$E = E \bigcup \{e \ | I_{bin}(e_x, e_y) = 1\}$
$L = L \bigcup \{l = G^i(e_x, e_y) \ | I_{bin}(e_x, e_y) = 1\}$
$N = N \bigcup \{\vec{n}^*_{x,y} \ | I_{bin}(e_x, e_y) = 1\}$
$//n_i^\star$ adjusted to point towards the ROI
end for
//Extract N_1 oriented patches using $n_i^\star$
$P = \{(p_i, l_i) \ | l_i \in \{1, 2, 3, 4, 5\}, i = 1...N_1\}$

Step1: Extract low-level features and train the 1st random forest classifier

$d_i \leftarrow \mathcal{F}(p_i)$ //low-level feature vector
$\{(d_i, l_i) \ | l_i \in \{1, 2, 3, 4, 5\}, i = 1...N_1\} \Longrightarrow RF_1$

{**Layer 2: Train the 2nd classification layer**}

Step2: Collect N_2 pairs of patches

$\{(e_i, l_i, n_i) | l_i \in \{0, 1, 2\} \wedge \angle n_i \in [0, \pi), i = 1...N_2\}$

0: a non-polyp edge,

1: a polyp edge where n_i points toward the polyp,

2: a polyp edge where -n_i points toward the polyp

Algorithm 2

for $i = 1...N_2$ //for each edge
  //Assume two normals
  $\{n_i^1 \leftarrow n_i, n_i^2 \leftarrow -n_i\}$
  p_i^1 oriented patch with n_i^1 being the normal
  p_i^2 oriented patch with n_i^2 being the normal
  Step3: Extract features
  $d_i^1 \leftarrow \mathcal{F}(p_i^1)$, $d_i^2 \leftarrow \mathcal{F}(p_i^2)$
  //Apply the first classifier RF_1
  $f_i^1 \leftarrow RF_1(d_i^1), f_i^2 \leftarrow RF_1(d_i^2)$
  Step4: Concatenate features
  $f_i \leftarrow c(f_i^1, f_i^2)$
end for
Step5: Train the 2nd random forest classifier
$\{(f_i, l_i)|l_i \in \{0, 1, 2\}, i = 1..., N_2\} \Longrightarrow RF_2$

APPENDIX C

3D RECONSTRUCTION PIPELINE

The pseudocode of the suggested 3D reconstruction pipeline.

Input:

- Camera parameters listed in Table 4.1
- Stream of tracking data
- Stream of depth information

Output:

- A point cloud P

3D reconstruction

{**3D projection at** $t = 1$**}**
  Acquire camera location and orientation from the tracking sensor
  Acquire a depth image from the depth camera
  Project pixels to the 3D space using Eq. 4.4 $\rightarrow P^1$
  Assign all the 3D points to the cloud $P = P^1$
  Assign an importance value of 1 to all the 3D points in the cloud $\lambda_i = 1$
{**3D projection at** $t > 1$**}**
  for $t = 2, 3, ..., N$ //for each edge
    Acquire camera location and orientation from the tracking sensor
    Acquire a depth image from the depth camera
    Project pixels to the 3D space using Eq. 4.4$\rightarrow P^t$
    Register P^t to the exiting cloud P using Eq. 4.5 $\rightarrow P_T^t$
    for $i = 1, 2, ..., |P_T^t|$ //for each point in P_T^t
      Get p_i, the i^{th} point from P_T^t
      Find the nearest neighbor point p_{nn} in the existing cloud to p_i
      Compute the distance to the nearest neighbor $d_i = \|p_{nn} - p_i\|$
      if $d_i > d_{MAX}$
        p_i is an outlier point and thus removed from P_T^t
      else if $d_i < d_{MIN}$
        p_i already exists in P and thus removed from P_T^t
        Increment the importance of p_{nn}, $\lambda(p_{nn}) = \lambda(p_{nn}) + 1$
      end if
    end for
  Add the P_T^t to the existing cloud, $P = P \cup P_T^t$
  Set the importance of the new points to one $\lambda_j = 1, j = 1, 2, ..., |P_T^t|$
end for