
STABILIZATION AND IMAGE LABELING FOR NBI ENDOSCOPIC IMAGE RECOGNITION

大腸NBI内視鏡画像認識の安定化と領域分割

By

TSUBASA HIRAKAWA



Department of Information Engineering
Graduate School of Engineering
HIROSHIMA UNIVERSITY

A thesis submitted to HIROSHIMA UNIVERSITY in accordance with
the requirements of the degree of DOCTOR OF PHILOSOPHY.

NOVEMBER 2017

AUTHOR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the requirements of the Hiroshima University and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

Tsubasa Hirakawa
November 2017

“Education never ends, Watson.

It is a series of lessons with the greatest for the last.”

– SHARLOCK HOLMES

in "The Adventure of the Red Circle" by Sir Arthur Conan Doyle

I would like to dedicate this thesis to my loving parents...

ACKNOWLEDGEMENTS

This dissertation would not have been impossible without the help and the guidance of several individuals around me, who in one way or another contributed and extended their valuable assistance in the preparation and completion of this work.

First and foremost, my greatest gratitude to my supervisor Dr. Toru Tamaki, whose sincerity and encouragement, I will never forget. Dr. Toru Tamaki has been my inspiration as I hurdle all the obstacles in the completion of my PhD life.

As well I express my application to Dr. Kazufumi Kaneda, Dr. Bisser Raytchev, and Dr. Takio Kurita of Hiroshima University whose support and comment greatly help me to conduct research.

My sincere gratitude to Dr. Laurent Najman of ESIEE Paris and Dr. Chaohui Wang of Université Paris-Est Marne-la-Vallée, whose advice were greatly helpful to conduct research. The work in Chapter 4 stems from an exciting collaboration with them. Their advice has been great help especially for image labeling. Also, I am really grateful to Dr. Yukiko Kenmochi of Université Paris-Est Marne-la-Vallée, who make an opportunity for the collaboration research.

I am grateful to Dr. Tetsushi Koide, Mr. Takumi Okamoto of Research Institute for Nanodevice and Bio Systems for their various opinions in menthly meeting. I immensely thank to Dr. Shigeto Yoshida of Hiroshima General Hospital of West Japan Railway Company and Dr. Shinji Tanaka, Dr. Yoko Kominami, and Dr. Rie Miyaki of Hiroshima University Hospital Endsocope Specialty for thier offer of endoscopic images and movies, without which my studies would not have been possible. The cooperative research with medical science brought very stimulating, precious, and critical experience to me.

And then, I thank to my laboratory members. Especially, Thanks to Mr. Shoji Sonoyama for supporting the real-time recognition system development.

Last but not the least, my family and my friends, for all your kindness understanding and supports. If there were no their supports, it would not have been possible to finish my study. Thank you very much.

This work was supported in part by JSPS KAKENHI grants numbers 14J00223, 26280015, and 24591026.

ABSTRACT

Colorectal endoscopy is widely used to diagnose colorectal cancer throughout the world and the recent development of narrow-band imaging system enables endoscopists to perform examinations in a short time. However, the intra/inter-observer variability show that the diagnosis can be subjective and highly depend on the endoscopist's experience. Hence, a computer-aided diagnosis system providing an objective measure for diagnosis could be an important diagnostic support during examinations. To this end, the prototype of the system has been developed, which recognize the center of video frame of endoscope and provide classification results to endoscopists in a frame by frame manner. However, this prototype system has two problems: one is that the output of the system is highly unstable because each frame is processed independently. The other is that the system recognizes only part of images or video frames.

In this thesis, we propose three methods to overcome these problems. The first method is a temporal smoothing method for posterior probability curves with a particle filter of Dirichlet distribution that introduces defocus information to likelihood estimation. The second method is an image labeling method with Markov random field and posterior probabilities obtained from a pretrained classifier. The third method is an image labeling method based on a tree of shapes and histogram features computed on the tree structure.

Keywords: Colorectal cancer, Endoscopy, Narrow-Band Imaging, Particle filter, Dirichlet distribution, Rayleigh distribution, Defocus information, Image labeling, Markov random fields, Texture analysis, Tree of shapes, Histogram feature, Mathematical morphology

TABLE OF CONTENTS

	Page
List of Tables	xiii
List of Figures	xv
1 Introduction	1
1.1 Colorectal Cancer	2
1.1.1 General Situation of Colorectal Cancer	2
1.1.2 Colorectal Endoscopy (Colonoscopy)	3
1.1.3 Medical Tumor Classification in Colorectal Endoscopy	5
1.2 Previous Work	8
1.2.1 NBI Patch Recognition System	9
1.2.2 NBI Videoendoscopy Recognition System	10
1.2.3 Problems of the Recognition System	11
1.3 Thesis Overview	12
2 Defocus-aware Dirichlet Particle Filter	15
2.1 Related Work	16
2.2 DPF	17
2.2.1 Particle Filters	17
2.2.2 Dirichlet Distribution	18
2.2.3 State Transition	18
2.2.4 Likelihood	20
2.3 D-DPF	21
2.3.1 Update Step	22
2.3.2 Hidden Variable γ_t	23
2.3.3 Prediction Step	26
2.3.4 Algorithm	27
2.4 Experimental Results	28
2.4.1 Dataset and Frame-wise Classification	28
2.4.2 Classification Results for Blurred Patches	29

TABLE OF CONTENTS

2.4.3	Results for Synthetic Video Sequences	30
2.4.4	Results on Real Endoscopic Videos	36
2.5	Summary	40
3	SVM-MRF image labeling of NBI Endoscopic Images	43
3.1	Related Work	43
3.2	SVM-MRF Image Labeling	44
3.3	Considering Highlight Regions	45
3.3.1	Detecting Highlight Regions	45
3.3.2	MRF Model Considering Highlight Regions	46
3.4	Experimental Results	46
3.4.1	Dataset and Evaluation Method	46
3.4.2	Labeling Results without Highlight Regions	47
3.4.3	Labeling Results with Highlight Regions	49
3.5	Summary	50
4	Tree-wise Discriminative Subtree Selection	53
4.1	Related Work	55
4.2	Tree of Shapes and SITA	56
4.2.1	Tree of Shapes	57
4.2.2	SITA	58
4.2.3	Recursive Representation of SITA	59
4.3	Proposed Method	59
4.3.1	Optimization	61
4.3.2	Labeling Procedure	62
4.4	Experimental Results	64
4.4.1	Energy Convergence	64
4.4.2	Labeling Results on a Synthetic Dataset	67
4.4.3	Labeling Results on the MSRC-21 Dataset	69
4.4.4	Labeling Results on the NBI Endoscopic Images	74
4.5	Summary	79
5	Conclusion	83
A	Development of Real-Time Classification System	85
A.1	Design of the Video Stream Capturing System	86
A.2	Implementation of an Endoscopic Video Frame Classification System	88
A.3	Results and Discussions	90
A.4	Summary	93

Bibliography	95
---------------------	-----------

LIST OF TABLES

TABLE	Page
4.1 Dice coefficients of labeling results on NBI endoscopic images.	79

LIST OF FIGURES

FIGURE	Page
1.1 The time trend of the deaths number cause of colorectal cancer in Japan.	2
1.2 5-year survival rate of colorectal cancer.	3
1.3 Colorectal endoscopy.	4
1.4 Images showing different colonoscopy processes.	5
1.5 Pit-pattern classification of colorectal lesions.	6
1.6 NBI magnification findings.	7
1.7 Overview of Bag-of-Visual Words.	8
1.8 Training sample construction by trimming a rectangle from an NBI videoendoscope image.	9
1.9 Performance of NBI patch recognition system.	10
1.10 Overview of real time recognition system.	11
1.11 Example of frame-wise classification results from an NBI video with snapshots. . . .	12
2.1 Example of defocus.	16
2.2 Examples of 3D Dirichlet distributions.	19
2.3 Examples of state transition probabilities modeled by Dirichlet distribution.	21
2.4 Examples of likelihood functions modeled by Dirichlet distribution.	22
2.5 Graphical models of DPF and D-DPF.	23
2.6 The concept of isolated pixel.	24
2.7 Histogram of IPRs computed from endoscopic videos.	25
2.8 The relationship between IPR, Rayleigh distribution and Dirichlet distribution. . . .	26
2.9 Proposed scaling function of $\sigma(z_t)$	27
2.10 Classification performance on the NBI image patches with and without Gaussian blur. 29	
2.11 Smoothing results on a synthetic video with Gaussian noise of standard deviation $\sigma_{noise} = 1$	31
2.12 Smoothing results for a synthetic video using Dirichlet noise with parameter $s = 20$. .	32
2.13 RMSE of the smoothing results for the synthetic videos.	34
2.14 Smoothing results for a synthetic video with Dirichlet noise with parameter $s = 50$ with three classes of type A, B, and C3.	35

2.15	Smoothing results for a synthetic video with Dirichlet noise with parameter $s = 50$ with three classes of type A, B, and C3 over the different number of particles.	36
2.16	Computational cost of D-DPF per frame.	37
2.17	Smoothing results on a real endoscopic video labeled as type B.	38
2.18	Smoothing results on real endoscopic videos labeled as types A and C.	39
3.1	Results of detecting highlights.	46
3.2	Evaluation procedure of labeling results.	47
3.3	Performance of labeling results for different values of p	48
3.4	Labeling result for NBI images of Type A.	49
3.5	Labeling result for NBI images of Type B.	49
3.6	Labeling result for NBI images of Type C3.	49
3.7	Labeling result for NBI images of Type A-1 for highlight regions.	50
3.8	Labeling result for NBI images of Type A-2 for highlight regions.	50
3.9	Labeling result for NBI images of Type B-1 for highlight regions.	50
3.10	Labeling result for NBI images of Type B-2 for highlight regions.	51
3.11	Labeling result for NBI images of Type C3-1 for highlight regions.	51
3.12	Labeling result for NBI images of Type C3-2 for highlight regions.	51
3.13	Labeling result for NBI images of Type C3-3 for highlight regions.	51
4.1	Overview of the proposed method.	55
4.2	Example of a synthetic image and corresponding tree of shapes.	57
4.3	Examples of labeling results using subtrees with different node sizes.	60
4.4	Example of a texture image sub-dataset.	64
4.5	Cost function over different initial values and thresholds for a training image.	65
4.6	Cost function values against the threshold over different iterations, starting with different initial values.	66
4.7	Energies and thresholds at each iteration.	67
4.8	Cost function values against the threshold over different iterations with different scale parameter values.	68
4.9	Labeling results on a synthetic sub-dataset.	69
4.10	Labeling results on a synthetic sub-dataset.	70
4.11	Labeling results with different scale parameter values λ	70
4.12	Labeling results on a synthetic sub-dataset.	71
4.13	Labeling results on a synthetic sub-dataset.	71
4.14	Labeling results on a synthetic sub-dataset.	72
4.15	Box plots for Dice coefficients over different numbers of training images.	72
4.16	Some failure labeling results on a synthetic sub-dataset.	73
4.17	Some failure labeling results on a synthetic sub-dataset.	74

4.18	Box plots for Dice coefficients over different w_{agg} on subset 2 and subset 9 of the MSRC-21 dataset.	75
4.19	Labeling results on a subset 2 of the MSRC-21 dataset.	76
4.20	Labeling results on a subset 9 of the MSRC-21 dataset.	77
4.21	Box plots for Dice coefficients over different numbers of training images on subset 2 and subset 9 of the MSRC-21 dataset.	78
4.22	Examples of images in the NBI endoscopic image dataset.	79
4.23	Labeling results on the NBI endoscopic images.	80
4.24	Some failure examples on the NBI endoscopic images.	81
A.1	The system configuration of our developed CAD system (Olympus EVIS LUCERA). .	86
A.2	The system configuration of our developed CAD system (Olympus EVIS LUCERA ELITE).	87
A.3	The endoscopic system and the developed system.	90
A.4	Screen shots of the developed video frame classification system.	91
A.5	Computational cost per frame.	92

INTRODUCTION

Current colonoscopy or colorectal endoscopy, an endoscopic examination of the colon, is capable of not only observing colorectal tumor but also treating with an endoscope. If a tumor would be in its early stage without suspicion of progress or metastasis of cancer, it might be possible to resect it during inspection. However, we need to understand the range and condition of the tumor correctly for treatment. This is because resecting a tumor might cause perforation, and resecting shallow might cause to stimulate proliferating and metastasizing capacity. Accordingly, endoscopists are required diagnosing precisely during colonoscopy. However, a small tumor tends to be overlooked visually because of difficulty of an operating endoscope and the shape of tumors. Due to chronic shortage of doctors, moreover, the number of patients per doctor and the burden of endoscopists are increasing annually.

In general, observations and treatments of tumors are done on different days even if those can be done simultaneously to enhance the accuracy of diagnosis. At the time of observation, endoscopists take a lot of pictures of a polyp, and later, other endoscopists decide the level of tumors using the pictures. Our research project provides a recognition system during examinations. According to the description of the level, endoscopists presume the existence and invasion depth of polyps. Because the decision depends on the experiences of endoscopists, it may be different for each endoscopist. Therefore, we provide a quantitative and statistical recognition measurement based on machine learning for endoscopists to reduce diagnostic error and inter/intra-observer variability. The recognition results of our system can be data to inform endoscopists of the possibility of diagnostic error, but not a correct decision.

To this end, we already proposed a narrow-band imaging (NBI) endoscopic patch recognition system with bag-of-visual words (BoVW) and a support vector machine (SVM) for a three-class classification problem based on the NBI magnification findings and achieved a high recognition

rate. Furthermore, we have developed a prototype system for NBI endoscopic video sequence, which is an extension of the above system to video sequences. However, there are two problems in the systems. One is that the output of the system is highly unstable because each frame is processed independently. The other is that the system recognizes only central parts of movies. This thesis proposes three methods to improve these problems. For the former, we propose a smoothing posterior probabilities with a particle filter of Dirichlet distribution. For the latter, we propose a segmentation method using Markov random field (MRF) framework and posterior probabilities obtained from SVM and another method based on a tree of shapes and histogram features computed on the tree structure.

1.1 Colorectal Cancer

This section describes fundamental knowledge of colorectal cancer and endoscopic examination such as statistical data, inspection, treatment, and microscopic classification to be endoscopist's diagnosis index.

1.1.1 General Situation of Colorectal Cancer

Colorectal cancer has been one of the most critical disease in the world. In Japan, the report estimates that 50,000 people have died from colorectal cancer in 2015 [108]. Figure 1.1 shows the trend of deaths of colorectal cancer in Japan. The number is increasing year by year, and have increased by 1.6 times over 20 years. And also, 49,190 people have died in 2016 in the United States [114], and 15,903 people have died in 2014 in the United Kingdom [15]. World

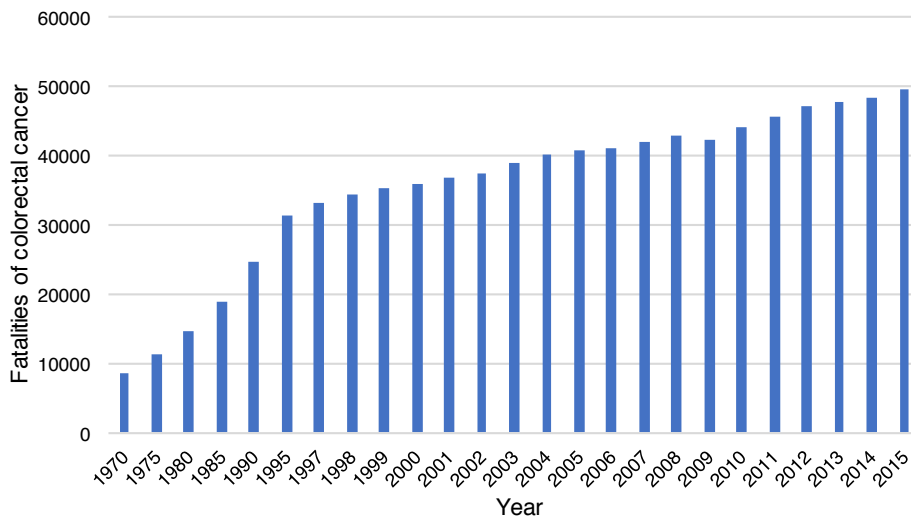


Figure 1.1: The time trend of the deaths number cause of colorectal cancer in Japan [108].

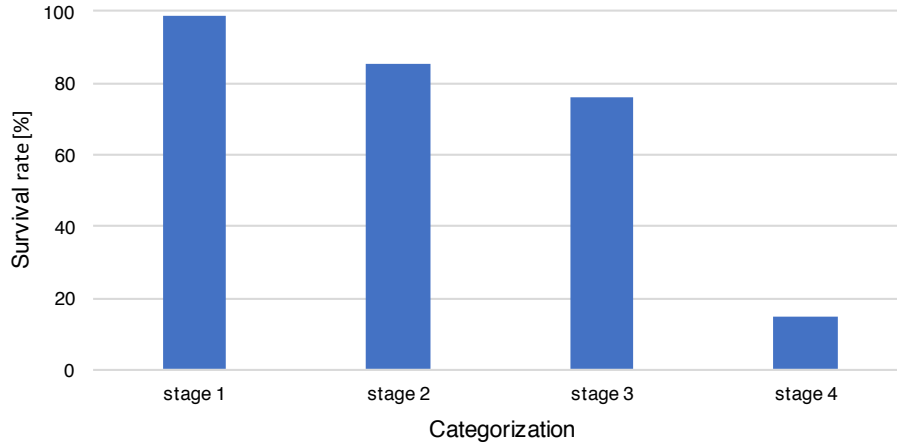


Figure 1.2: 5-year survival rate of colorectal cancer [43].

Health Organization has released projections [56] in which the number of deaths in the world is estimated to be about 780,000 in 2015, and is expected to rise to 950,000 in 2030.

The 5-year survival rate [43] is a medical barometer which measures the prognosis of disease (Figure 1.2). The rate indicates the ratio of the patient who survives after progress for five years from diagnosis. Colorectal cancer is classified into four stages (stage 1 ~ 4) based on the progress. In stage 1, the cancer has grown into submucosa and has not spread to nearby lymph or distant sites. Then, the cancer has grown into the muscularis propria, a deeper, thick layer of muscle that contracts to force the contents of the intestines along in stage 2. The cancer has spread into the lymph nodes in stage 3. Furthermore, the cancer has attached to other organs and structures. The percentages of late stage is less than 20, while for early stage nearly 100. Therefore, it is important to detect a cancer in its early stage. This clarify that colorectal cancer can be curable if the cancer is detected in early stage. However, patients which colorectal cancer often have no subjective symptoms, and cancer might be already developed when the symptoms occur. Therefore, it is desirable to be inspected on a regular basis.

1.1.2 Colorectal Endoscopy (Colonoscopy)

The inspection methods of colorectal cancer include Fecal Occult Blood Test (FOBT) [57, 132], Digital Rectal Examination (DRE) [39], biomedical markers [73], CT colonography [54, 160], MR colonography [3, 136], Marvin Positron Emission Tomography (PET) [91], Ultrasound [123], Double Contrast Barium Enema (DCBE) [16, 68], Confocal Laser Endomicroscopy (CLE) [75], and Virtual Endoscopy [122]. Among them, *colorectal endoscopy* (or *colonoscopy*), an inspection to observe and treat a colon inserting a slender pipe with a small charge-coupled device (CCD) camera or lens, is widely used in hospitals as a standard medical procedure. Figure 1.3 shows the process of colorectal endoscopy. In colorectal endoscopy, a scope of endoscopy is inserted into near the end of the small intestine firstly, and endoscopists then observe the inside of the colon

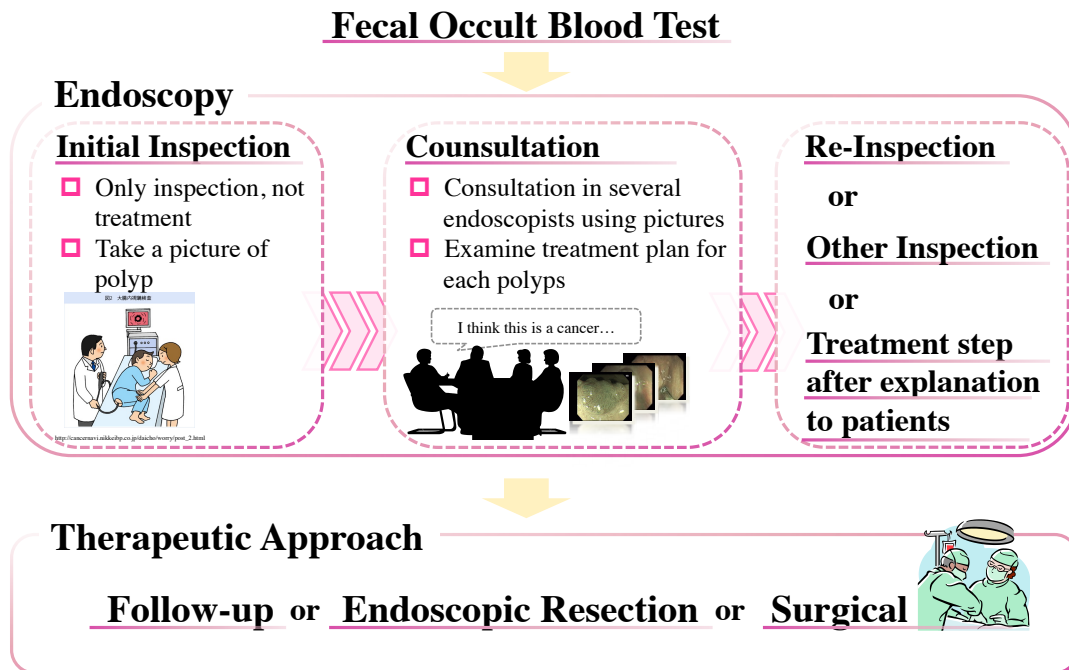


Figure 1.3: Colorectal endoscopy.

with returning the scope. The lesions of colorectal tumors on the colon surface are often visually inspected. Endoscopists take a lot of pictures, and resect the polyp in some cases. However, there are some parts to be difficult to observe such as the reverse side of the folds of the intestinal wall and an incipient tumor has a small lesion. Therefore, endoscopists might overlook polyps. To enable endoscopists to examine more precisely during examination, endoscopic devices have been developed. Zoom-videoendoscope with a magnification factor of up to 100 [147] is one of such developed endoscopic device. Because endoscopists diagnose the presence of colorectal cancer and the invasion depth by microstructures of tissue, zoom-videoendoscope enables to observe in detail.

To diagnose histologically by using the visual appearance of colorectal tumors in endoscopic images, *chromoendoscopy* is used. During chromoendoscopy, indigo carmine dye spraying or crystal violet staining are used to enhance the microscopic appearances of the pit patterns illuminated by a white light source. Figure 1.4a shows an image of a colon taken by an endoscope without staining, while Figures 1.4b and 1.4c show images stained by two different dyes. In 1.4b and c, the structure of the mucosal surface on the polyp is well enhanced and the visibility is much better than in white light colonoscopy 1.4a.

Narrow-band imaging (NBI) [37, 38, 100, 133] is recently developed videoendoscopic system that uses RGB rotary filters placed in front of a white light source to narrow the bandwidth of the spectral transmittance. The central wavelength of the RGB filters are set to 540 and 415 nm with a bandwidth of 30 nm, because the hemoglobin in the blood absorbs lights of these

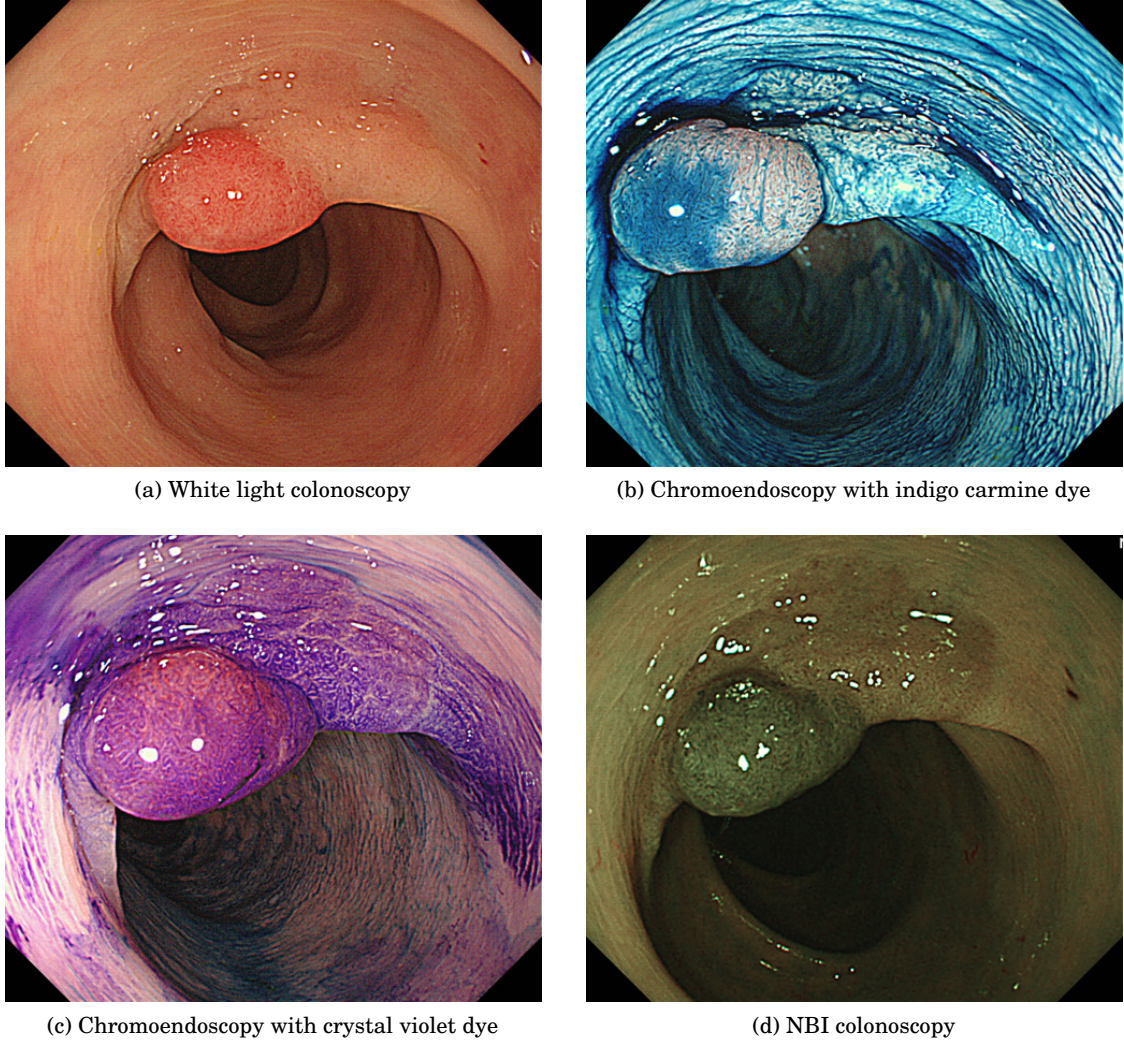


Figure 1.4: Images showing different colonoscopy processes.

wavelengths. NBI provides a limited penetration of light to the mucosal surface, and enhances the micro vessels and their fine structure on the colorectal surface (see Figure 1.4d). NBI enables an endoscopist to quickly switch a white light colonoscopy image to an NBI colonoscopy image when examining tumors, while a chromoendoscopy requires a cost for spraying, washing and vacuuming dye and water. In this thesis, we use images and video sequences of colorectal cancers taken by NBI colonoscopy.

1.1.3 Medical Tumor Classification in Colorectal Endoscopy

As described in the preceding section, endoscopists diagnose a polyp by using the visual appearance. In the past clinical research, the microscopic appearances enhanced by chromoendoscopy are visually inspected, which is called *pit-pattern classification*. In recent years, classification

using NBI system has been standard, which is called *NBI magnification findings*. Herein, we describe these two type of visual assessment strategies.

1.1.3.1 Pit-pattern Classification

A *pit pattern* is the shape of a *pit* [2], the opening of a colorectal crypt, and can be used for the visual inspection of mucosal surface. The microscopic appearances of the pit patterns are enhanced by chromoendoscopy, which enables endoscopists to classify progress level of tumor. This is called the *pit-pattern classification* and is a standard visual inspection method of colonoscopy.

Pit-pattern analysis started in the 1970s [2, 79], and developed over the next two decades [67, 81, 82]. Figure 1.5 shows the most widely used pit-pattern classification, which categorizes pit-pattern into types I to V. Types III and V are further divided into III_S (S: Smaller) and III_L (L: Larger), V_I (I: Irregular) and V_N (N: Non-structure), respectively.

The pit-pattern classification has been used to differentiate non-neoplastic colorectal lesions from neoplastic ones, and to guide therapeutic decisions. Indicated diagnosis roughly corresponds to: follow up (no resection) (type I, and II), endoscopic resection (type III_S, III_L, and IV), surgery (type V_N), and further examinations (type V_I).

I		Round pit (normal pit)			
II		Asteroid pit			
III _S		Tubular or round pit that is smaller than the normal pit (Type I)			
III _L		Tubular or round pit that is smaller than the normal pit (Type I)			
IV		Dendritic or gyrus-like pit			
V _I		Irregular arrangement and sizes of III _S , III _L , IV type pit pattern			
V _N		Loss or decrease of pits with an amorphous structure			

Figure 1.5: Pit-pattern classification of colorectal lesions [144].

1.1.3.2 NBI Magnification Findings

NBI has been introduced to overall digestive organs such as gastro, esophageal, and colorectal examinations from the early 2000's. Especially, gastro and esophageal examinations demands NBI due to the irritation of dye-spraying. NBI has been used for colorectal examination from around 2004 [37, 38, 100, 133]. Because of the high visibility of the microvessels [18, 60, 66], NBI is used for pit-pattern analysis [63] and microvessel analysis [62].

NBI magnification findings is the view aiming at objective and qualitative diagnosis of colorectal cancer by evaluating pit-pattern classification and microvessel structures synthetically. Several categorizations of NBI magnification findings have been developed by different medical research groups.

- *Hiroshima University Hospital*: three main types (A, B, and C) and subtypes (C1, C2, and C3).
- *Sano Hospital*: three main types (I, II, and III) and subtypes (IIIA and IIIB) based on the density of capillary vessels, lack of uniformity, ending and branching of the vessels.
- *Showa University Northern Yokohama Hospital*: six main types (A to F) associated with two subtypes (1 and 2) based on the thickness, network structure, density and sparsity of the vessels.
- *The Jikei University School of Medicine*: four main types (1, 2, 3 and 4) and subtypes (3V and 3I) based on detail and regularity of the vessels.

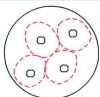
Type A		Microvessels are not observed or extremely opaque.				
Type B		Fine microvessels are observed around pits, and clear pits can be observed via the nest of microvessels.				
Type C	1		Microvessels comprise an irregular network, pits observed via the microvessels are slightly non-distinct, and vessel diameter or distribution is homogeneous.			
	2		Microvessels comprise an irregular network, pits observed via the microvessels are irregular, and vessel diameter or distribution is heterogeneous.			
	3		Pits via the microvessels are invisible, irregular vessel diameter is thick, or the vessel distribution is heterogeneous, and a vascular areas are observed.			

Figure 1.6: NBI magnification findings [71].

Among them, in this thesis we use the classification proposed by Tanaka's group at Hiroshima University Hospital. Figure 1.6 shows NBI magnification findings proposed by Tanaka's group. It divides the microvessel structure in an NBI image into type A, B, C (see Figure 1.6). Type C is divided into three subtypes C1, C2, and C3 according to detailed texture.

Colonoscopy, however, is affected by the skill and familiarity of each endoscopist, and the burden of endoscopists is increasing due to the increased number of patients. Moreover, inter/intra-observer variability [106, 107, 119] shows that diagnosis can be subjective and depends on the endoscopist's experience. Therefore, it is important to develop a computer-aided systems able to provide supporting diagnosis for this type of cancer [143] and a number of studies have been conducted.

1.2 Previous Work

We have proposed an NBI patch recognition system and extended to recognition of NBI video sequences as our prior work. In this section, we describe about the details of these systems, the performance, and the problems.

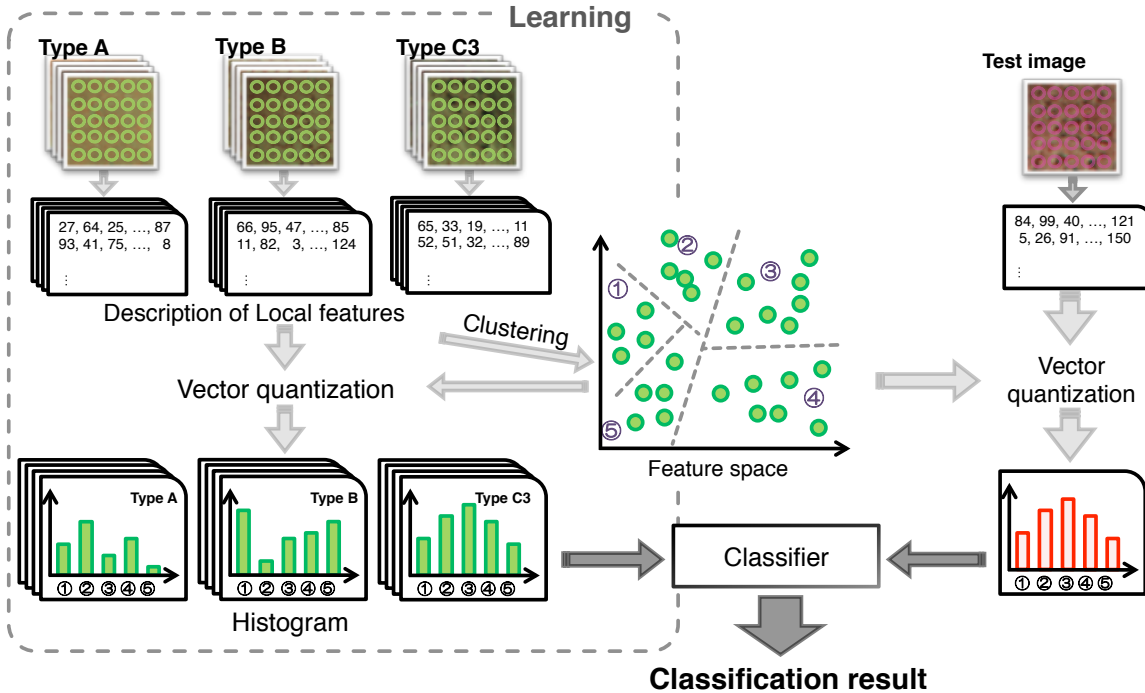


Figure 1.7: Overview of BoVW.

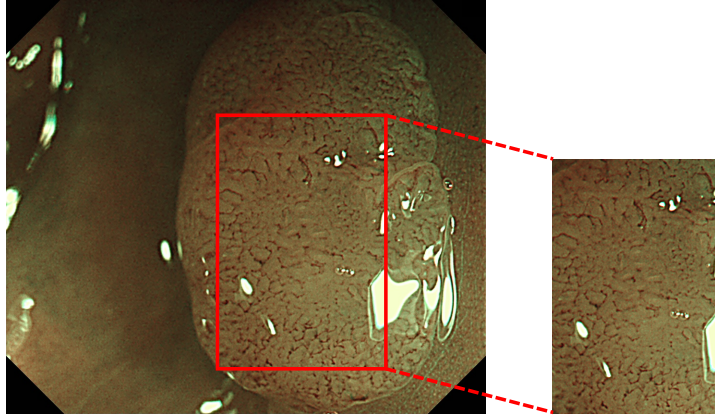


Figure 1.8: Training sample construction by trimming a rectangle (right) from an NBI videoendoscope image (left).

1.2.1 NBI Patch Recognition System

Tamaki et al. [145, 146] have proposed an NBI patch recognition system with BoVW [22, 88, 118] for a three-class classification problem (types A, B, and C3) based on the NBI magnification findings. The overview of the BoVW is shown in Figure 1.7. BoVW represents an image as a histogram of representative local features extracted from the image regardless of their location. As a local features, they used densely sampled scale-invariant feature transform (SIFT) [31, 59, 70, 98, 99]. They used a descriptor of each different scale at a sample point as an element of single feature vector: for example, if there are 100 points on the grid and 4 scales are being used, then we have 400 descriptors called gridSIFT descriptors, where each descriptor is a 128 dimensional vector. An alternative way to define a descriptor is to combine the descriptors of different scales at the sample point into a single feature vector: e.g., in this case we have 100 descriptors, each being a 512 dimensional vector (we call this *variant multi-scale gridSIFT*).

To create visual words histogram, they used two methods: hierarchical k-means clustering [116] and class-wise concatenation of visual words [161]. In their experiments, they need to cluster several millions of features densely sampled from almost one thousand training images, and explore vocabularies of size up to several thousands visual words. Hence, it is necessary to use hierarchical k-means for reducing the computational cost. Because the scaling is well-known to affect the classification performance, each bin of the histogram was scaled linearly so that the range of values of the bin was $[-1, 1]$ throughout for the training samples; we then stored the scaling factor to scale histograms of test data.

In a classifier, they used an SVM [21, 134, 142, 153] with five kernel types (radial basis function (RBF), linear, χ^2 , χ^2_{ng} , and histogram intersection (HI)). During training phase, an SVM was trained with the visual word histogram features of the training samples for different values of the penalty parameter using a 5-fold cross validation. The values that gave the best performance were selected and stored.

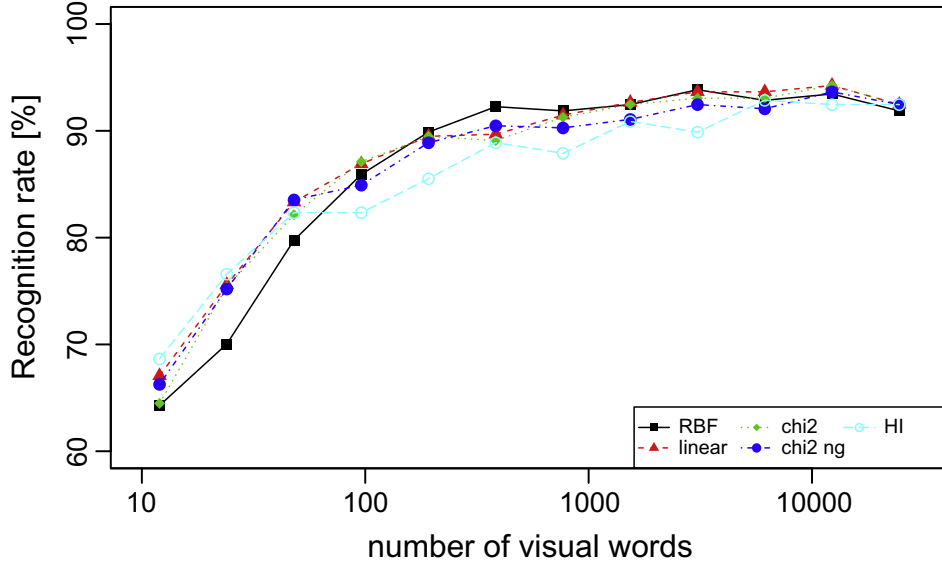


Figure 1.9: Performance of NBI patch recognition system with different kernels (quoted from [145]). As parameters of gridSIFT, grid spacing is set to 5 pixels and scale is set to 5 and 7 pixels. Used SVM kernels are RBF, linear, χ^2 (chi2), χ_{ng}^2 (chi2 ng), and HI.

To train the SVM classifier, we collected training samples as follows. As still image of a video frame was trimmed (see Figure 1.8) by endoscopists into a rectangular patch that contained typical microvessel structures. Then, labels were assigned to the image patches by endoscopists. Note that all these trimmed images were collected and all experiments of our previous works and this thesis were performed at the Hiroshima University Hospital. The guidelines of the Hiroshima University ethics committee were followed, and informed consent was obtained from the patients and their families. We are collecting such image patches to construct large-scale dataset and we have 2247 NBI image patches (Type A: 504, Type B: 847, Type C1: 257, Type C2: 57, Type C3: 582). Note that the number of used NBI image patches differ depending on the experiment because the NBI image patches are increasing year by year.

Figure 1.9 shows the performance of the NBI image recognition system. The recognition rate of 93 percent was obtained for 10-fold cross validation on 908 NBI image patches (Type A: 359, Type B: 462, Type C3: 87).

1.2.2 NBI Videoendoscopy Recognition System

The NBI patch recognition system enabled us to develop a real-time recognition and assessment system during endoscopic examination, which is expected to be of a great help for colonoscopy [143]. One straightforward way is *frame-wise classification*, or feeding an NBI videoendoscopic image sequence frame by frame to the recognition system: a rectangular window at the center of each frame is classified, and the output at each frame can be either a class label or the posterior probabilities of each class (showing the confidence in the estimated class labels).

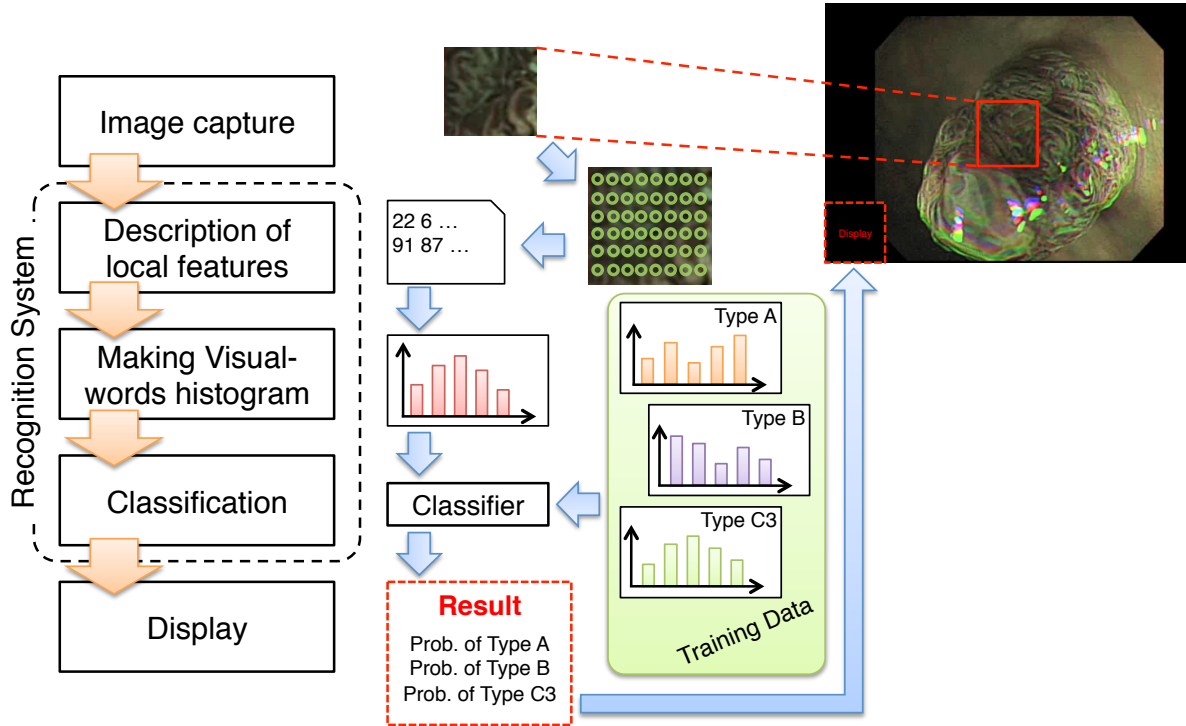


Figure 1.10: Overview of real time recognition system.

Figure 1.10 shows an overview of the real-time recognition system. The region-of-interest (ROI) is set to a rectangular patch at the center of the frame of the videoendoscope (a white rectangle inside the video frame in the upper right of Figure 1.10). Then, densely sampled SIFT descriptors [31, 59, 70, 98, 99] computed on a regular grid of 5 pixels with two different scales (5 and 7 pixels) are extracted in the ROI. The extracted SIFT descriptors are represented as a histogram of the visual words (representative SIFT features) computed by hierarchical k-means clustering [116]. Each bin of this histogram is linearly scaled with a fixed factor mentioned above. This visual word histogram feature is then classified by a pretrained linear SVM classifier [21, 134, 142, 153] to obtain the classification probabilities for each category. These results (probabilities) are finally displayed on a monitor by superimposing the results onto the video frame.

The classifier-training phase was conducted offline before the online classification. We trained a linear SVM classifier using the training samples mentioned above, and all the necessary parameters of the SVM classifier were stored in a file. During the online classification phase, this file was loaded and used for the SVM classification (therefore, we call it a pretrained classifier).

1.2.3 Problems of the Recognition System

The NBI videoendoscopy recognition system could be greatly helpful for endoscopists during examinations. However, this system has two critical problems.

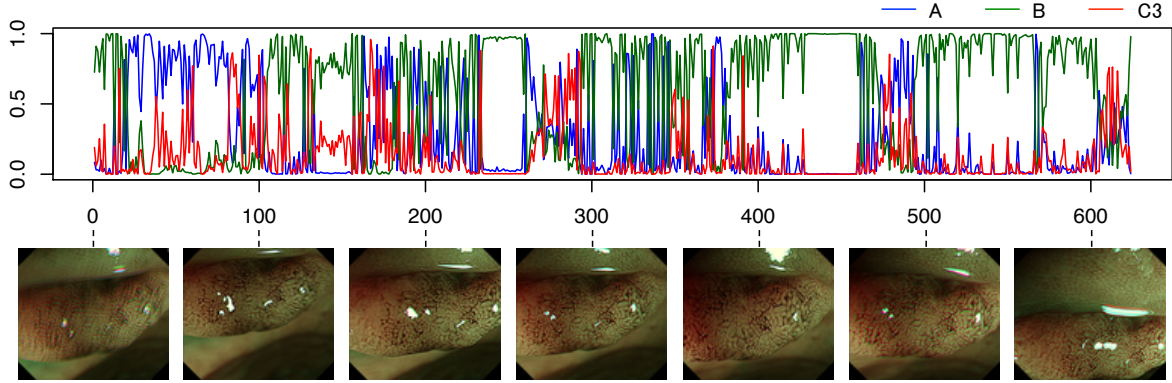


Figure 1.11: Example of frame-wise classification results from an NBI video [145] with snapshots. For each frame, a patch of size 200×200 at the center of the frame is classified by a frame-wise classifier to obtain posterior probabilities as a result of a 3-class classification problem. These three classes, type A, B, and C3, correspond to certain diagnostic criteria for a tumor. In the upper row, the three curves of posterior probabilities represent the classification results obtained in each frame: the horizontal axis shows frame number and the vertical axis the classification probabilities for the three classes of type A (blue), B (green), and C3 (red). The bottom row shows frames of the video at every 100 frames. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Firstly, this strategy is not appropriate for observations by endoscopists. Figure 1.11 shows a typical result obtained from a frame-wise classification with three classes. The three curves of posterior probabilities represent the classification results of every frame and are shown to visualize the confidence of the classifier. Although this video sequence continues to capture the same tumor, the classification results are highly unstable, and it would not be difficult for endoscopists to understand the output during an examination. The output therefore should be temporally smoothed and stabilized in order for endoscopists to easily understand the status of tumors of interest.

The second problem lies in the fact that they can only a part of images of the video frame. For instance, in case that a tumor is not in the center of the frame or multiple tumors exist in the frame, these systems cannot provide appropriate objective measures. Therefore, recognizing an entire endoscopic image would be a further assistance for endoscopists during examinations, and could be used to train inexperienced endoscopists.

1.3 Thesis Overview

This thesis proposes three methods to improve the aforementioned two problems.

For the first problem, that is instability of temporal classification results, we propose a method to smooth posterior probability curves temporally, that is based on a particle filter with Dirichlet distribution because multi-dimensional probability vector obtained from a classifier should be

smoothed. Moreover, the instability of classification results might be caused by incorrect feature vectors described in highly defocused video frames. Hence, we introduce defocus information as a confidence of such defocused video frame and use the defocus information to estimate a likelihood.

For the second problem, that is recognizing a part of a whole image, we deal with the problem as an image labeling problem and propose two methods. The first image labeling method is based on MRF framework with a posterior probabilities obtained from a classifier. Since we already have achieved higher classification performance on patch classification experiments, we simply extend this patch classification system to a whole endoscopic image. Posterior probabilities obtained from a classifier are used for a data term of MRF model. Then, the defined MRF model is optimized by α - β swap graph cut, and we obtain labeling results. In an endoscopic image, moreover, there exists highlight regions, which could affect to feature descriptor and provide wrong labeling results. Therefore, we consider highlight regions in a MRF model.

The second image labeling method is based on a tree of shapes and histogram features computed on the tree structure. In this method, we compute histogram features at every node in a tree of shapes. Image labeling can be done by classifying optimal subtrees in a tree of shapes which are discriminative and useful for labeling. The parameters to select such discriminative subtrees and a classifier are estimated by minimizing a cost function with training images.

The body of this report consists of five chapters, the first one is this introductory chapter. Chapter 2 proposes a smoothing method for posterior probability curves and demonstrates the experimental results. Chapter 3 proposes a image labeling method using MRF framework and posterior probabilities obtained from an SVM classifier. Chapter 4 tackles the image labeling problem again. We proposed a novel image labeling method based on a tree of shapes and histogram features computed on the tree structure. Finally, we give conclusions and discuss future work in Chapter 5.

DEFOCUS-AWARE DIRICHLET PARTICLE FILTER

As we introduced in Chapter 1, we propose a method to smooth and stabilize posterior probability curves. A computer-aided system that provides an objective measure for diagnosis to endoscopists during colonoscopy would be of great assistance [143]. To this end, we have developed frame-wise classification system, i.e., using a machine-learning-based classifier trained off-line with training image patches to recognize a part of every endoscopic video frame and showing classification results (labels or probabilities) on a screen. However, the problem then arises that we do not see when we independently classify training image patches. Figure 1.11 shows a typical result obtained from a frame-wise classification with three classes. The three curves of posterior probabilities represent the classification results of every frame and are shown to visualize the confidence of the classifier. Although this video sequence continues to capture the same tumor, the classification results are highly unstable, and it would be difficult for endoscopists to understand the output during an examination.

One of the principal causes for this instability is scene blur, or defocus, due to the narrow depth of field (see Figure 2.1). Since operating an endoscope requires expert skill, and the intestinal wall continues moving, it is difficult to maintain focus on a tumor for a long time. Features extracted from defocused frames cause unstable results because the classifier has not been trained with such features. Our preliminary experiments also demonstrate that classifying defocused image patches performs worse than classifying well-focused ones (see Section 2.4.2). Removing defocused frames from a video stream [120] would not be helpful in such an application because results on the screen would frequently stop or disappear.

To overcome this problem, we propose a method for smoothing probability curves, or sequences of posterior probabilities, such as those shown in Figure 1.11 on the basis of a particle filtering with a Dirichlet distribution, which is called the *Dirichlet particle filter* (DPF). Furthermore, we

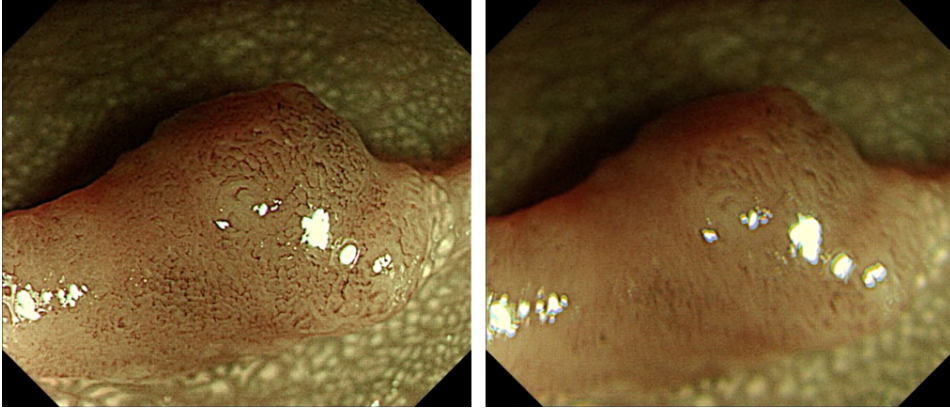


Figure 2.1: Example of defocus. The tumor is captured in focus in one frame (left), but is defocused in another frame (right).

incorporate information representing the degree of defocus from each frame in the framework of DPF. We call this proposed method the *defocus-aware Dirichlet particle filter* (D-DPF). There are two reasons why we need to develop our own method for smoothing probability curves.

First, smoothing techniques use only given signals; therefore, it is difficult to recover from failures in frame-wise classification owing to the defocus of frames. In such a case, it is reasonable to use additional information which represents defocus of each frame; smoothing results tend to follow the observation of the current frame if the frame is in focus, and to keep the results from the previous frames otherwise. Our proposed method uses isolated pixel ratio (IPR) (see Section 2.3.2.1) as defocus information in the likelihood of the particle filter so as to show the confidence of the classification result at each frame.

Second, smoothed results obtained by existing smoothing methods must typically be renormalized at each frame to sum to one, leading to inconsistency between frames as this has no probabilistic significance. Our system outputs confidence values at each frame, i.e., posterior probabilities for the results when classifying a patch in each frame into three classes (type A, B, and C3), on the basis of NBI magnification findings [71, 119] (see Section 1.1.3.2). Therefore, we developed a probabilistic framework with a particle filter to perform “smoothing of probabilities” using the Dirichlet distribution (see Section 3.2) in such a way that defocus information is incorporated.

2.1 Related Work

Polyp detection has been the most widely performed and studied task in colorectal videoendoscopy in the past two decades. Maroulis et al. [105] proposed a detection system of colorectal lesions in endoscopic videos using neural networks, and Karkanis et al. [72] used color wavelet features. There have also been various other efforts [8, 65, 90, 124].

Surprisingly, classification of endoscopic videos has been scarcely investigated. One possible

reason might be that a frame-wise classification could be developed by simply applying patch-based classification to video streams frame by frame. In fact, many patch-based classification methods for endoscopic images have been proposed for pit-pattern [44–53, 83–85] and NBI-endoscopic images [42, 141, 145, 151]. Such a simple application of frame-wise classification to video frames was proposed by Manivannan et al. [103]. They classified video frames into normal and abnormal using patch statistics and Gaussian scale-space.

Later, they proposed a *video-specific SVM*, training with video frames to independently classify images or frames [104]. This approach involves the following problems. First, each video frame must have a label assigned by endoscopists, which is a very expensive task. Second, an endoscopic video frame contains many unnecessary parts such as dark background, defocused parts, and highlights. Therefore, using entire frames for learning would lead to a deterioration of classification performance. Third, training a classifier with an entire video is more expensive than training with image patches. Selecting representative image patches is much more efficient when training a classifier or constructing a training dataset. Hence, we employ a more practical strategy, Åsmoothing as post-processing of a frame-wise classification.

Several methods have been proposed to detect and exclude defocused frames from endoscopic videos. Oh et al. [120] attempted to classify video frames into informative and non-informative ones. They proposed two methods, i.e., edge- and clustering-based methods. As an edge-based method, they apply a Canny edge detector to each frame and calculate the IPR, the ratio of isolated edge pixels to all edge pixels. They then classify each frame by thresholding the IPR. As a clustering-based method, they extract seven texture features from gray-level co-occurrence matrices of discrete Fourier transform magnitude images and then classify each frame by k-means clustering. To detect indistinct frames, Arnold et al. [1] used the L^2 -norm of the detail coefficients of a wavelet conversion. Liu et al. [96] proposed robust tracking by detecting and discarding endoscopic video frames that lack features useful to track by using the blurry image detection algorithm proposed by Liu et al. [95].

In our method, we use Oh’s IPR because it is inexpensive to compute and incorporate into a particle filter.

2.2 DPF

In this section, we develop DPF, which is a particle filter with a Dirichlet distribution.

2.2.1 Particle Filters

Particle filtering [14, 40, 76, 130] is a method for estimating the internal states of a state-space model sequentially. In particle filters, a probability distribution is represented by Monte Carlo approximation, i.e., by a set of K random samples, or particles, to handle non-linear and non-Gaussian problems. A particle filter estimates posterior probabilities of the internal states

conditioned on observations through the following two steps. A prediction step estimates the prior probabilities at time t from observations before time $t - 1$ and a state transition. An update step estimates the posterior probabilities at time t from the prior probabilities and a likelihood. To avoid confusion of terms, we hereafter refer to *classification probability* as the discrete (posterior) probability obtained from a frame-wise classification, and to *posterior probability* as the smoothed probability obtained by the particle filter.

Observation $\mathbf{y}_t$ is the classification probability obtained at time (or frame) t and it has a unit L_1 norm; $\|\mathbf{y}_t\|_1 = 1$. Let $\mathbf{y}_{1:t} = \{\mathbf{y}_1, \dots, \mathbf{y}_t\}$ be a series of observations obtained prior to time t . The internal state $\mathbf{x}_t$ is the posterior probability that should be estimated at time t by smoothing. Hence $\|\mathbf{x}_t\|_1 = 1$, and we use the notation $\mathbf{x}_{1:t} = \{\mathbf{x}_1, \dots, \mathbf{x}_t\}$ in the same manner as that for $\mathbf{y}_{1:t}$.

A prediction step estimates the prior probability $p(\mathbf{x}_t | \mathbf{y}_{1:t-1})$ using the following integral:

$$(2.1) \quad p(\mathbf{x}_t | \mathbf{y}_{1:t-1}) = \int p(\mathbf{x}_t | \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1} | \mathbf{y}_{1:t-1}) d\mathbf{x}_{t-1},$$

where $p(\mathbf{x}_t | \mathbf{x}_{t-1})$ is a *state transition probability* between states at time $t - 1$ and t . Once $p(\mathbf{x}_t | \mathbf{y}_{1:t-1})$ is computed, an update step computes the posterior probability $p(\mathbf{x}_t | \mathbf{y}_{1:t})$ as follows:

$$(2.2) \quad p(\mathbf{x}_t | \mathbf{y}_{1:t}) \propto p(\mathbf{y}_t | \mathbf{x}_t) p(\mathbf{x}_t | \mathbf{y}_{1:t-1}),$$

where $p(\mathbf{y}_t | \mathbf{x}_t)$ is the *likelihood*. Repeating the prediction and update steps from time zero yields a sequence of estimates of $\mathbf{x}_t$.

2.2.2 Dirichlet Distribution

To represent probability distributions of $\mathbf{x}_t$, we propose to use the N -dimensional Dirichlet distribution [7] with parameter $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)$, $\alpha_i > 0$ defined by

$$(2.3) \quad \text{Dir}_{\mathbf{x}}[\boldsymbol{\alpha}] = \frac{\Gamma(\sum_{i=1}^N \alpha_i)}{\prod_{i=1}^N \Gamma(\alpha_i)} \prod_{i=1}^N x_i^{\alpha_i - 1},$$

where Γ is the gamma function and $\mathbf{x} = (x_1, \dots, x_N)$ is a random variable with $\|\mathbf{x}\|_1 = 1$. The parameter $\boldsymbol{\alpha}$ controls the shape of distribution. Figure 2.2 shows examples of three-dimensional (3D) Dirichlet distributions. When $\alpha_i < 1$ for all i , the density has greater probabilities around the vertices as shown in figure 2.2(c). When $\alpha_i = 1$ for all i , the density flattens. Otherwise, the density has a peak at the mode $\frac{\boldsymbol{\alpha} - \mathbf{1}}{\|\boldsymbol{\alpha}\|_1 - N}$, where $\mathbf{1}$ is a vector of ones, as shown in Figure 2.2(e). Additionally, the larger the parameter values are, the steeper the peak becomes. In our three-class classification problem ($N = 3$), the support of the probability density is a two-dimensional triangle in a 3D space. Using the Dirichlet distribution enables us to formulate the state transition and the likelihood of the model.

2.2.3 State Transition

A prediction step models the relationship between internal states at time $t - 1$ and t . To define a state transition probability, we remember that the internal state of our problem is a posterior

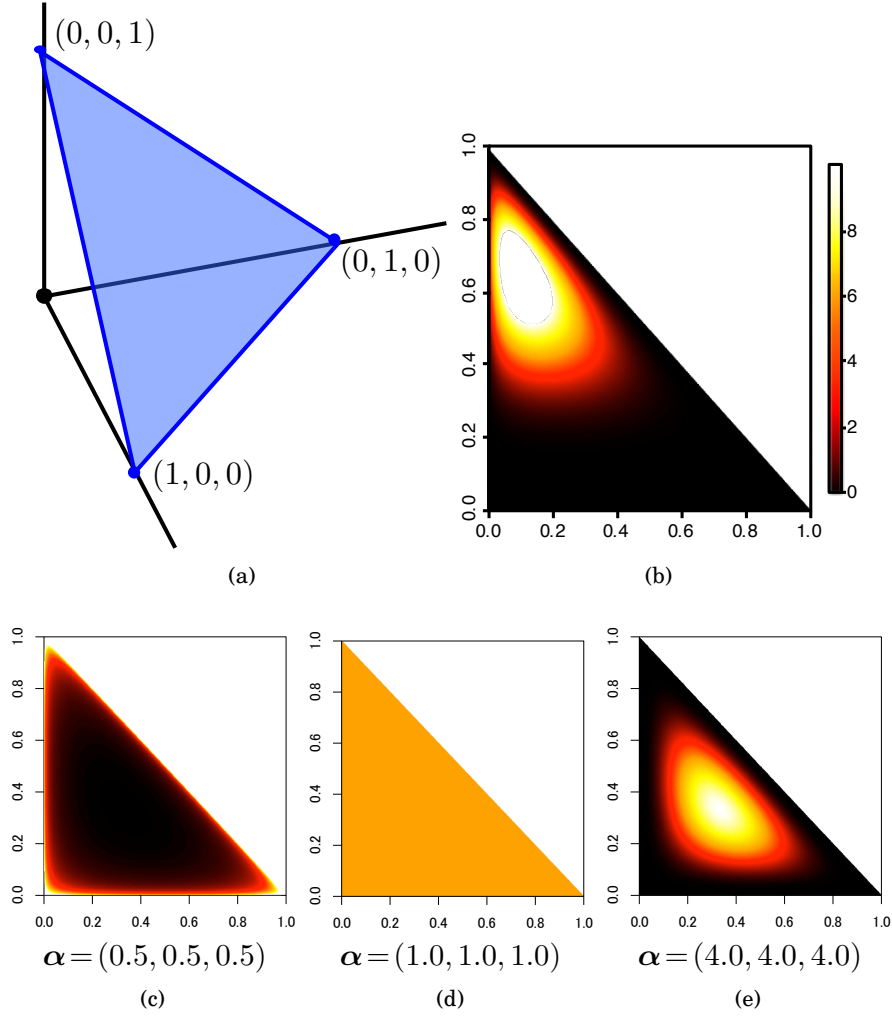


Figure 2.2: Examples of 3D Dirichlet distributions. (a) Support of the Dirichlet distribution. (b) Probability density of a Dirichlet distribution (darker the pixels, lower the density). (c, d, and e) Typical probability density shapes for different parameters α .

probability $\mathbf{x}$ that satisfies $\|\mathbf{x}\|_1 = 1$. The support of $\mathbf{x}$ is exactly the same as that of the Dirichlet distribution; thus, it can be used to define the state transition. Intuitively, internal state $\mathbf{x}_t$ has a large probability if it is similar to $\mathbf{x}_{t-1}$. Therefore, it is natural to use the probability density $p(\mathbf{x}_t | \mathbf{x}_{t-1})$ having a peak around $\mathbf{x}_{t-1}$. We propose to define the state transition probability with a Dirichlet distribution as follows:

$$(2.4) \quad p(\mathbf{x}_t | \mathbf{x}_{t-1}) = \text{Dir}_{\mathbf{x}_t}[\alpha(\mathbf{x}_{t-1})],$$

where parameter α is now a function of $\mathbf{x}_{t-1}$. To constrain the density (2.4) to have a peak around $\mathbf{x}_{t-1}$, we assume a linearity between α and $\mathbf{x}_{t-1}$:

$$(2.5) \quad \alpha = A\mathbf{x}_{t-1} + \mathbf{b},$$

where A is an $N \times N$ matrix and $\mathbf{b}$ is an N -vector. Throughout the remainder of this chapter, we further simplify the linear function as $A = aI$ and $\mathbf{b} = b\mathbf{1}$, and use the following simplified notation:

$$(2.6) \quad \alpha(\mathbf{x}_{t-1}, a, b) = a\mathbf{x}_{t-1} + b\mathbf{1}.$$

Now we redefine Eq. (2.4) as follows:

$$(2.7) \quad p(\mathbf{x}_t | \mathbf{x}_{t-1}, \theta) = \text{Dir}_{\mathbf{x}_t}[\alpha(\mathbf{x}_{t-1}, \theta, 0)].$$

Here we set $b = 0$ to make the mean of the density coincide with $\mathbf{x}_{t-1}$:

$$(2.8) \quad E[\mathbf{x}_t] = \frac{\alpha}{\|\alpha\|_1} = \frac{a\mathbf{x}_{t-1}}{\|a\mathbf{x}_{t-1}\|_1} = \frac{a\mathbf{x}_{t-1}}{a\|\mathbf{x}_{t-1}\|_1} = \mathbf{x}_{t-1},$$

where we use the scale-invariant property of the $L1$ -norm for the third equation and $\|\mathbf{x}_{t-1}\|_1 = 1$ for the last equation. Figure 2.3 shows examples of the state transition probability density function of a 3D Dirichlet distribution. According to our observations when changing the range of the parameter θ , 100 or greater is a typical choice for the value of θ .

2.2.4 Likelihood

In an update step, particles representing prior distribution $p(\mathbf{x}_t | \mathbf{y}_{1:t-1})$ are weighted by a likelihood, and the posterior probability $p(\mathbf{x}_t | \mathbf{y}_{1:t})$ is then estimated on the basis of the weighted particles. In our problem, we assume that the likelihood has a peak at $\mathbf{y}_t$ and propose to define the likelihood by using Dirichlet distribution as follows:

$$(2.9) \quad p(\mathbf{y}_t | \mathbf{x}_t, \gamma) = \text{Dir}_{\mathbf{x}_t}[\alpha(\mathbf{y}_t, \gamma, b)].$$

Here, we use the probability distribution of $\mathbf{x}_t$ as the likelihood of $\mathbf{y}_t$. We wish to make the likelihood have a broad peak at $\mathbf{y}_t$ because the smoothing effect might decrease if the peak sufficiently steep such that only particles near $\mathbf{y}_t$ have extremely large weights. To this end, we set $b = 1$ instead of $b = 0$, because the zero bias leads to a steep peak when the value of $\mathbf{y}_t$ is extremely close to 0. Figure 2.4 shows examples with and without the bias term (i.e., $b = 0$ or $b = 1$). When some values in $\mathbf{y}_t$ are negligible as in Figure 2.4(b) and (c), the likelihood with $b = 0$ has a steep peak at the edge of the triangular support. In contrast, the likelihood with $b = 1$ has a reasonably broad peak inside the triangle. Based on our observations, typical values of γ and b should be 10 (or less) and 1 (or greater), respectively. Particularly, setting $b = 1$ makes the likelihood have a peak at exactly $\mathbf{y}_t$ because:

$$(2.10) \quad \begin{aligned} \text{mode}[\mathbf{x}_t] &= \frac{\alpha - \mathbf{1}}{\|\alpha\|_1 - N} \\ &= \frac{(\gamma\mathbf{y}_t + \mathbf{1}) - \mathbf{1}}{\|\gamma\mathbf{y}_t + \mathbf{1}\|_1 - N} \\ &= \frac{\gamma\mathbf{y}_t}{\gamma\|\mathbf{y}_t\|_1 + \|\mathbf{1}\|_1 - N} \\ &= \mathbf{y}_t, \end{aligned}$$

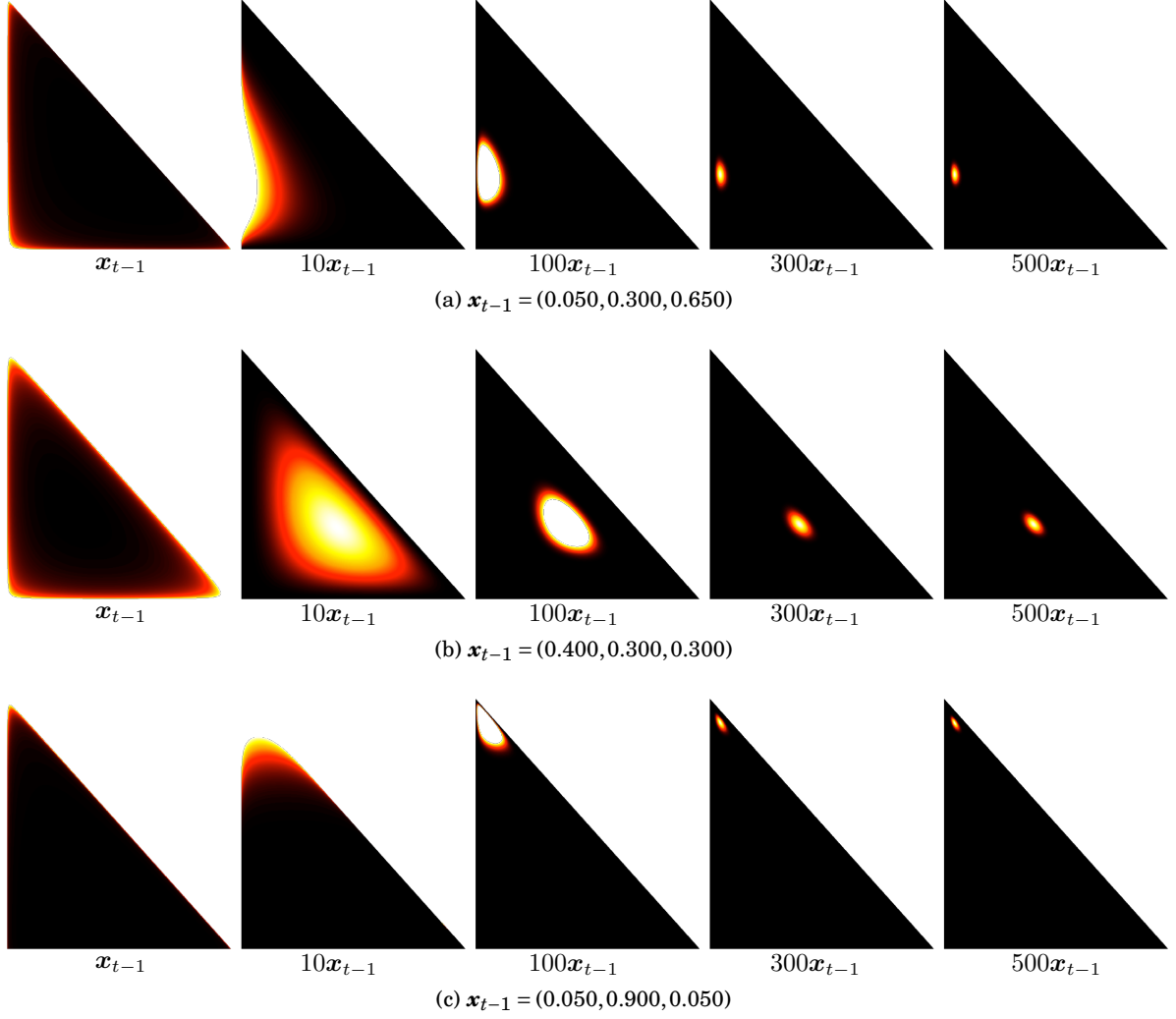


Figure 2.3: Examples of state transition probabilities modeled by Dirichlet distribution.

where $\|\mathbf{y}_t\|_1 = 1$ and $\|\mathbf{1}\|_1 = N$.

2.3 D-DPF

Here, we incorporate additional defocus information at each frame to develop D-DPF. The DPF discussed in the previous section assumes that each observation $\mathbf{y}_t$ is generated by the true state $\mathbf{x}_t$ with a Dirichlet distribution with parameter γ . The graphical model of DPF is shown in Figure 2.5(a), with factor nodes (black squares) representing potential functions of $\mathbf{y}_t, \mathbf{x}_t$, and the deterministic parameter γ .

We assume that observation $\mathbf{y}_t$ is influenced by the true state $\mathbf{x}_t$ as well as a temporal hidden variable γ_t , which is inferred from the additional defocus information. The graphical model of

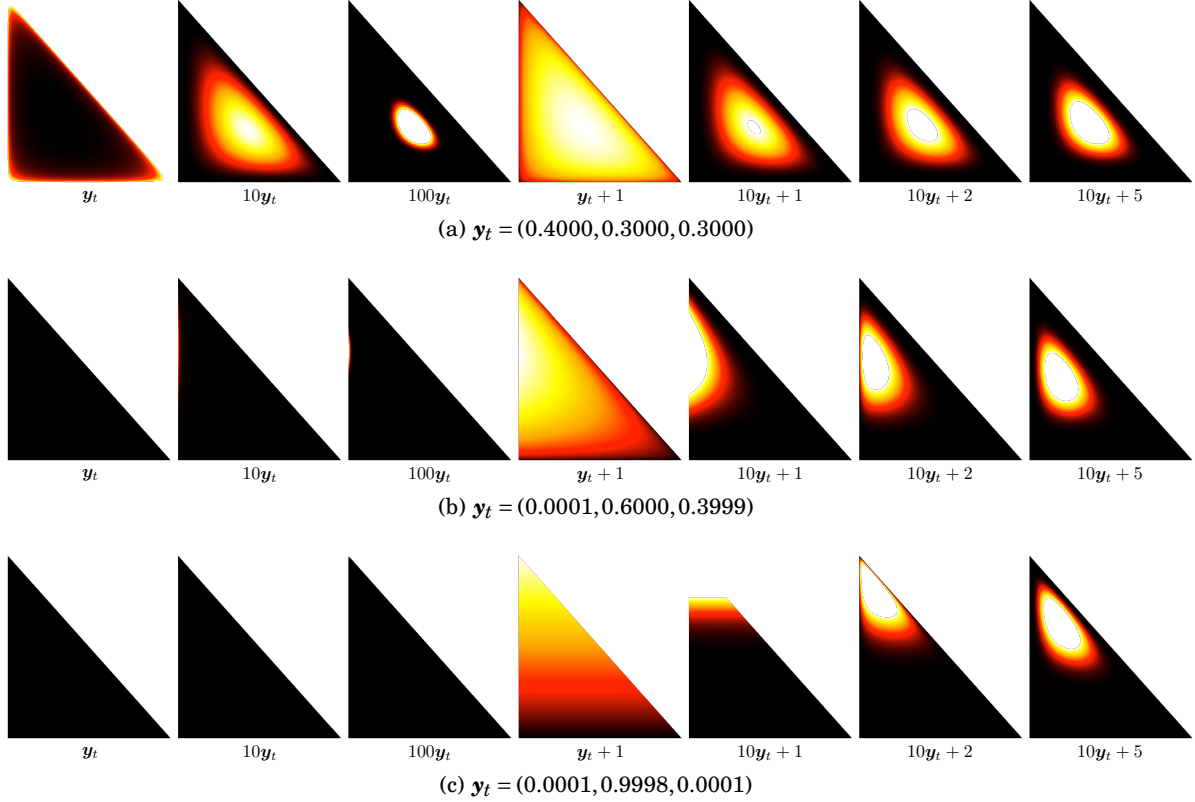


Figure 2.4: Examples of likelihood functions modeled by Dirichlet distribution.

D-DPF in this section is shown in figure 2.5(b). The factor nodes now represent $\mathbf{y}_t, \mathbf{x}_t$, and γ_t . We further assume that the hidden variable is generated by the defocus information z_t .

2.3.1 Update Step

We begin with a modified definition of the update step as follows:

$$(2.11) \quad p(\mathbf{x}_t | \mathbf{y}_{1:t}, \gamma_{1:t}, z_{1:t}) \propto p(\mathbf{y}_t, \gamma_t, z_t | \mathbf{x}_t) p(\mathbf{x}_t | \mathbf{y}_{1:t-1}, \gamma_{1:t-1}, z_{1:t-1}),$$

where z_t is a scalar value representing the defocus information at time t .

We model the likelihood $p(\mathbf{y}_t, \gamma_t, z_t | \mathbf{x}_t)$ with a Dirichlet distribution with a peak at $\mathbf{y}_t$, whose broadness depends on z_t . According to the graphical model in figure 2.5(b), we propose to redefine the likelihood as follows:

$$(2.12) \quad \begin{aligned} p(\mathbf{y}_t, \gamma_t, z_t | \mathbf{x}_t) &= p(\mathbf{y}_t, \gamma_t | \mathbf{x}_t) p(z_t | \mathbf{x}_t, \mathbf{y}_t, \gamma_t) \\ &= p(\mathbf{y}_t, \gamma_t | \mathbf{x}_t) p(z_t | \gamma_t), \end{aligned}$$

Here, we use the fact that z_t is conditionally independent of $\mathbf{y}_t$ and $\mathbf{x}_t$ given γ_t based on the graphical model. The potential function $p(\mathbf{y}_t, \gamma_t | \mathbf{x}_t)$, corresponding to the factor node in the

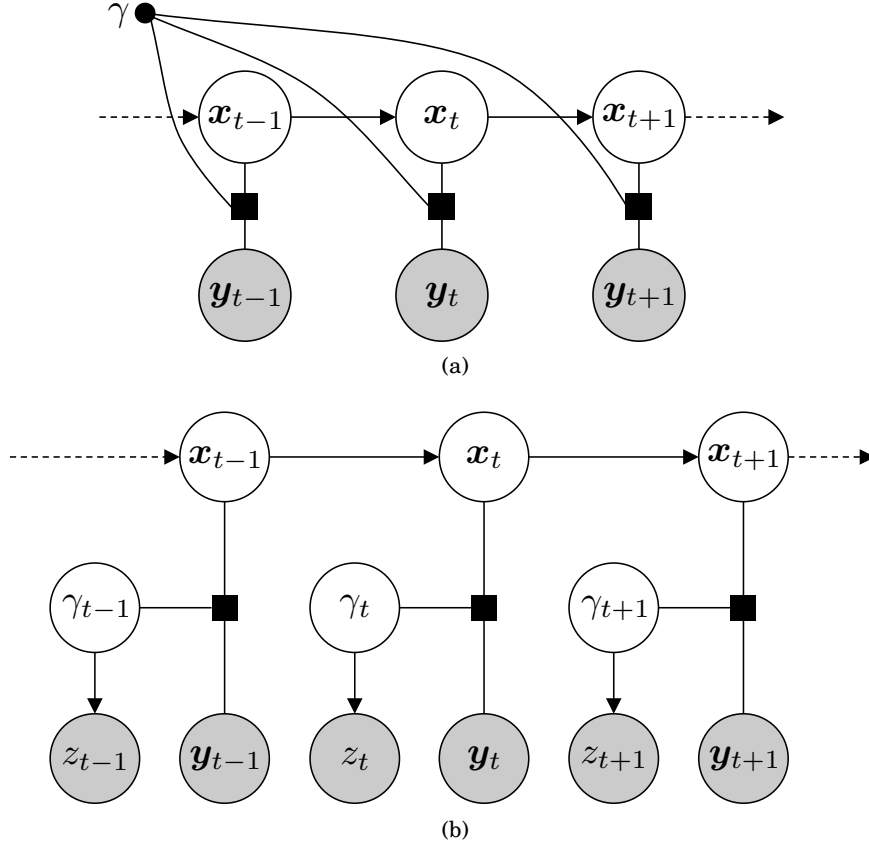


Figure 2.5: Graphical models of (a) DPF and (b) D-DPF.

graphical model, has a form similar to Eq. (2.9); hence, we define it as follows:

$$(2.13) \quad p(\mathbf{y}_t, \gamma_t | \mathbf{x}_t) = \text{Dir}_{\mathbf{x}_t}[\boldsymbol{\alpha}(\mathbf{y}_t, \gamma_t, 1)].$$

2.3.2 Hidden Variable γ_t

We model $p(z_t | \gamma_t)$, the relation between the hidden variable γ_t , and the defocus information z_t . We wish to reduce the effect of classification failures of a frame-wise classifier at defocused frames. In that case, observation $\mathbf{y}_t$ is less reliable, and the likelihood is expected to have a broad peak with the result that particles far from the peak at $\mathbf{y}_t$ are assigned larger weights. Therefore, we control γ_t to be smaller at defocused frames to have a broad likelihood. Herein, we use the IPR as z_t , and model $p(z_t | \gamma_t)$ with the Rayleigh distribution.

2.3.2.1 IPR

Isolated pixels are edge pixels extracted by a Canny edge detector, whose eight-neighbors are not edge pixels, as shown in Figure 2.6. IPR is the ratio of the isolated pixels to all edge pixels and takes values in the range between 0 and 1. We observe connected edge pixels in a sharp and

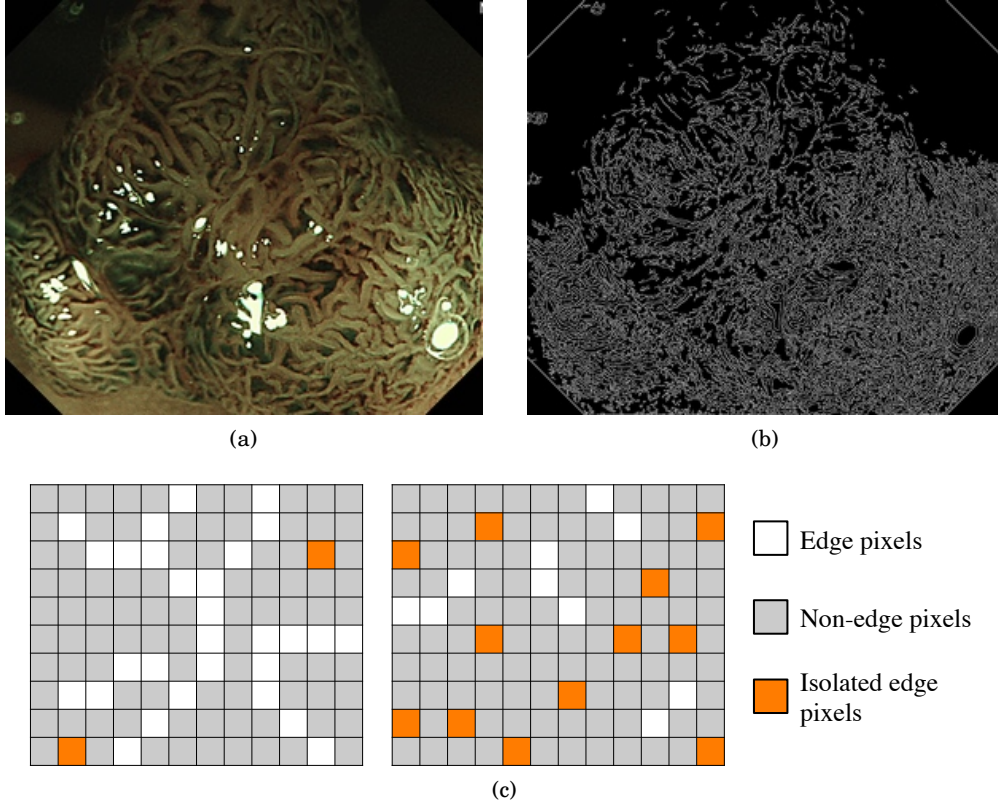


Figure 2.6: The concept of isolated pixel proposed by [120]. (a) An example of endoscopic image and (b) edges extracted by Canny edge detector. (c) Edges of focused (left) and defocused (right) frames.

focused image, whereas many isolated pixels are observed in defocused frames. In other words, a focused frame has a lower IPR value, and a defocused frame has a higher IPR value. IPR can be used to classify frames as informative or non-informative.

In Oh et al.’s paper, IPR values are distributed in the range between 0 and 0.1. However, observations can differ in different endoscopic videos due to frame size, zooming, and optical magnification, or when different types of endoscopes are used. To estimate the distribution of IPR in our endoscopic videos, we computed a histogram of IPR extracted from 33 videos (see Section 2.4.1 for details), as shown in Figure 2.7. We can see that IPR is distributed between 0 and 0.01. Using the IPR as z_t , we propose the model described in the next section.

2.3.2.2 Rayleigh Distribution

We use the Rayleigh distribution [27] to represent the relationship between the hidden variable γ_t and defocus information z_t . The Rayleigh distribution is defined by

$$(2.14) \quad \text{Ray}_\gamma[\sigma] = \frac{\gamma}{\sigma^2} \exp\left(-\frac{\gamma^2}{2\sigma^2}\right),$$

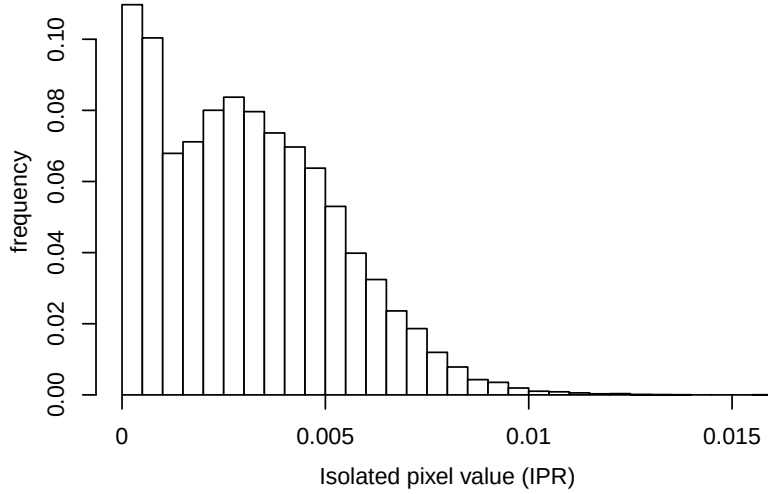


Figure 2.7: Histogram of IPRs computed from endoscopic videos.

where $\sigma > 0$ is a parameter. The top row of Figure 2.8 shows a few examples of the probability density function of the Rayleigh distribution. Smaller values of σ cause the distribution to peak toward zero, whereas larger values of σ broaden it.

As discussed above, we wish to have a broad peak of the Dirichlet distribution as a likelihood when the frame is defocused, and a lower value of γ_t is preferred in that case. In terms of IPR, a defocused frame contains many isolated pixels, resulting in larger IPR values. In summary, at a defocused frame, IPR or z_t is larger, and smaller values of γ_t must be sampled during the sampling procedure of the particle filter, resulting in a broad peak of the likelihood. Consequently, we propose to use z_t for controlling the parameter σ of the Rayleigh distribution as follows:

$$(2.15) \quad p(z_t | \gamma_t) = \text{Ray}_{\gamma_t}[\sigma(z_t)],$$

where $\sigma(z_t)$ is now a function of z_t . To achieve the desired behavior, we use a function of the form:

$$(2.16) \quad \sigma(z_t) = a \exp(bz_t),$$

where a and b are parameters to be tuned. A plot of this function is shown in Figure 2.9. The reason for using exponential decay is the range of z_t . If a frame is in focus, then the IPR might be zero or some small positive value. However, it can be extremely large (as much as one) for a defocused frame. Therefore, we assume that the range of z_t is $[0, 0.01]$, as mentioned above, but also allow larger values if they have little effect. The use of exponential decay allows larger values beyond the range above, but they would be effectively squeezed into an extremely narrow range on the vertical axis, as shown in figure 2.9.

As a reasonable range for the vertical axis, σ , we choose from 1 to 4 for σ , which is in accordance with observations of typical values of γ_t . At the end of the previous section, we mentioned that we prefer γ_t to take values of 10 or less. When we observe the horizontal axis γ of

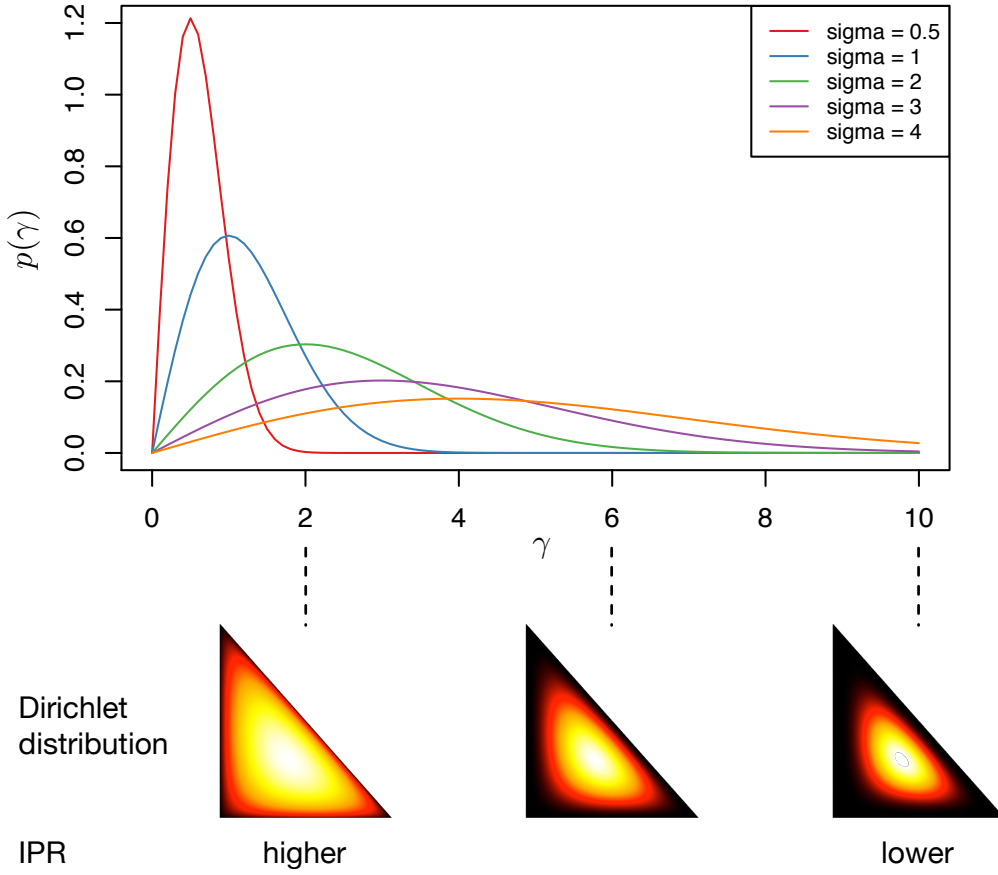


Figure 2.8: The relationship between IPR, Rayleigh distribution and Dirichlet distribution. Top: examples of the probability density function of the Rayleigh distribution. Bottom: Examples of Dirichlet distributions corresponding to different values in the Rayleigh distributions.

figure 2.8, the Rayleigh distribution with $\sigma = 4$ has support that almost covers the range $[0, 10]$. Therefore, in the current work we use the fixed (but flexible) range $[0, 0.01]$ for z_t , $[1, 4]$ for σ , and $[0, 10]$ for γ_t . To this end, we solved the following system of equations

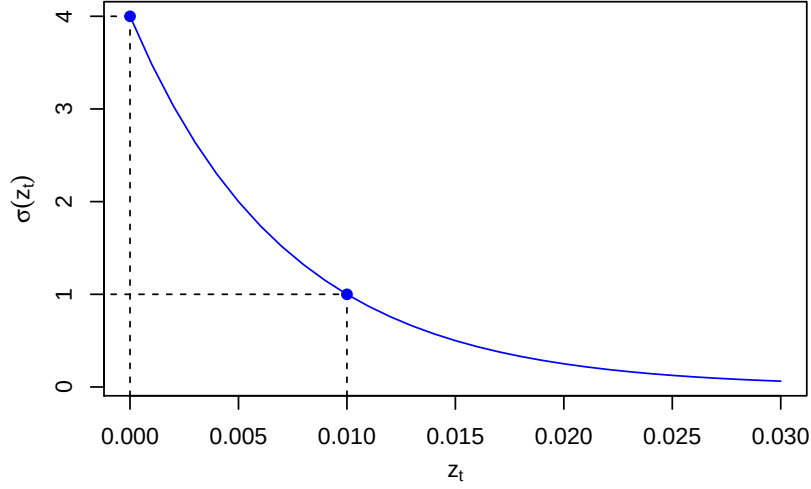
$$(2.17) \quad \begin{cases} 4 = a \exp(b \cdot 0) \\ 1 = a \exp(b \cdot 0.01), \end{cases}$$

to obtain $a = 4$ and $b = \frac{1}{0.01} \ln(\frac{1}{4}) = -\frac{\ln 4}{0.01}$.

2.3.3 Prediction Step

We use the same state transition as that discussed in Section 2.2 and define the prediction step as

$$(2.18) \quad p(\mathbf{x}_y | \mathbf{y}_{1:t-1}, \gamma_{1:t-1}, z_{1:t-1}) = \int p(\mathbf{x}_t | \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1} | \mathbf{y}_{1:t-1}, \gamma_{1:t-1}, z_{1:t-1}) d\mathbf{x}_{t-1},$$

Figure 2.9: Proposed scaling function of $\sigma(z_t)$.**Algorithm 1** Defocus-aware Dirichlet Particle Filter (D-DPF).

-
- 1: Sample K particles $\{\mathbf{s}_{0|0}^{(i)}\}_{i=1}^K$ from $p(\mathbf{x}_0)$.
 - 2: **for** time $t = 1 \dots T$ **do**
 - 3: **for** $i = 1 \dots K$ **do**
 - 4: Draw a sample $\mathbf{s}_{t|t-1}^{(i)} \sim \text{Dir}_{\mathbf{x}_t}[\boldsymbol{\alpha}(\mathbf{s}_{t-1|t-1}^{(i)}, \theta, 0)]$.
 - 5: **end for**
 - 6: Compute z_t from video frame at time t .
 - 7: Compute $\gamma_t \sim \text{Ray}_{\gamma_t}[z_t]$.
 - 8: **for** $i = 1 \dots K$ **do**
 - 9: Compute a weight $\pi_t^{(i)} = \text{Dir}_{\mathbf{x}_t = \mathbf{s}_{t|t-1}^{(i)}}[\boldsymbol{\alpha}(\mathbf{y}_t, \gamma_t, 1)]$.
 - 10: **end for**
 - 11: Sample K times as $\{\mathbf{s}_{t|t}^{(i)}\}_{i=1}^K$ from $\{\mathbf{s}_{t|t-1}^{(i)}\}_{i=1}^K$ with replacement according to the weights $\pi_t^{(i)}$.
 - 12: Estimate $\boldsymbol{\alpha}$ from $\{\mathbf{s}_{t|t}^{(i)}\}_{i=1}^K$.
 - 13: Compute the mode $\hat{\mathbf{x}}_t$ of the Dirichlet distribution from $\boldsymbol{\alpha}$.
 - 14: **end for**
-

where

$$(2.19) \quad p(\mathbf{x}_t | \mathbf{x}_{t-1}) = \text{Dir}_{\mathbf{x}_t}[\boldsymbol{\alpha}(\mathbf{x}_{t-1}, \theta, 0)].$$

2.3.4 Algorithm

Algorithm 1 details the proposed D-DPF. At each time step t , the mode $\hat{\mathbf{x}}_t$ of the Dirichlet distribution is obtained to visualize a plot along with the input observation $\mathbf{y}_t$ in the experiments:

1. Sample K particles according to an initial Dirichlet distribution $p(\mathbf{x}_0)$ with $\boldsymbol{\alpha} = (0.4, 0.3, 0.3)$ (line 1).

2. Sample prediction particles by performing transition of particles at time $t - 1$ according to the state transition probability (lines 2 to 5).
3. Compute the defocus information z_t and sample γ_t according to z_t (lines 6 and 7).
4. Estimate weights $\pi_t^{(i)}$ for prediction particles (lines 8 to 10).
5. Sample with replacement for $\mathbf{s}_{t|t}^{(n)}$ to be proportional to weight $\pi_t^{(i)}$ (line 11). Subsequently, compute the maximum likelihood estimate of α of Dirichlet distribution [113] from particles $\mathbf{s}_{t|t}^{(n)}$ (line 12). Then, each component of the mode $\hat{\mathbf{x}}_t$ of Dirichlet distribution is computed separately by

$$(2.20) \quad x_i = \frac{\alpha_i - 1}{\sum_{i=1}^N \alpha_i - N}, \alpha_i > 1.$$

In case that α_i is smaller than 1, we set $\alpha_i = 0$ (line 13).

6. Return to step 2.

2.4 Experimental Results

This section demonstrates the effectiveness of the proposed D-DPF smoothing. The following subsections describe the dataset of image patches and videos, and the classification results for blurred image patches and endoscopic video sequences.

2.4.1 Dataset and Frame-wise Classification

We used a dataset of 1671 NBI image patches of different sizes (type A: 504, type B: 847, type C3: 320) to train an SVM classifier for frame-wise classification. Each of the image patches was trimmed from an endoscopic video frame and labeled by endoscopists. These endoscopic video frames were captured and collected during endoscopic examinations, and each frame has the same label as the corresponding image patch. Details about the dataset, features used, and classification can be found in [145].

For evaluation, we have 33 NBI-endoscopic videos (type A: 5, type B: 27, type C3: 1) whose frame rate is 30 fps and size is full HD (1980×1080 pixels), wherein the window size displaying the endoscopic video is 1000×870 pixels. Each endoscopic video shows a single tumor, but there are many defocused frames. For each video frame, a 200×200 patch at the center of the window is classified by a pretrained frame-wise classifier to obtain classification probabilities $\mathbf{y}_t$. Frame lengths of the videos range between 200 and 2500, with more than 20,000 frames in total.

Labeling each video frame of these videos is, therefore, very expensive, as stated previously. Labeling still images in the aforementioned datasets of 1671 images was possible because it took several years to collect that number of NBI images for various patients. In Section 2.4.3, we instead use synthetic endoscopic video sequences, wherein frame labels are known, for analysis

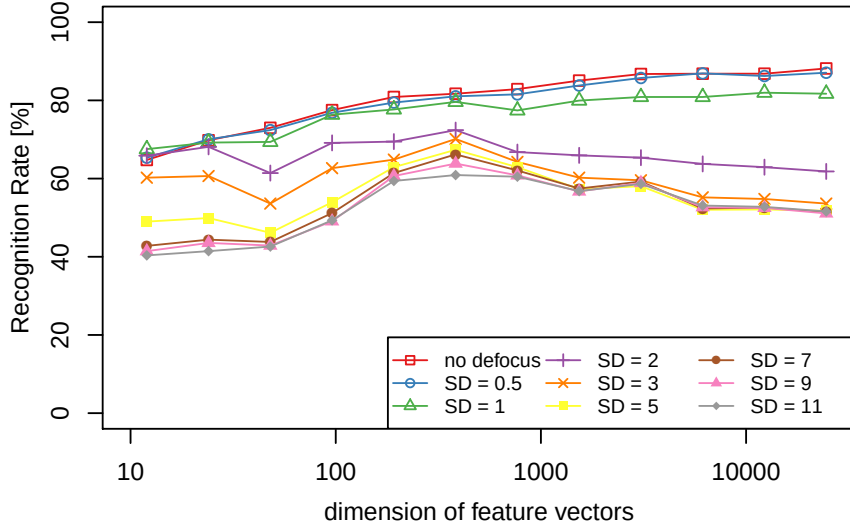


Figure 2.10: Classification performance on the NBI image patches with and without Gaussian blur. SD stands for σ_{blur} . The horizontal axis shows the dimension of the feature vectors (the number of visual words, see [145] for details).

and evaluation of the proposed method. In Section 2.4.4, we show some of the results for real videos to demonstrate the behavior of the proposed method.

The training NBI images have been used for clinical reports, while the endoscopic videos were collected for our experiments. All of these endoscopic images and videos were collected at the Hiroshima University Hospital, following the guidelines of the Hiroshima University ethics committee, and informed consent has been obtained from the patients and their families.

2.4.2 Classification Results for Blurred Patches

Before showing the results for the D-DPF, we demonstrate the performance deterioration of blurred image patch classification when using the classification method proposed by Tamaki et al. [145]. For training, 160 NBI image patches for each class, 480 NBI image patches in total, were randomly selected from the 1671 image patch dataset. The remainder of the dataset was used for evaluation by adding Gaussian blur with standard deviation $\sigma_{blur} = 0.5, 1, 2, 3, 5, 7, 9, 11$. The classification results are shown in figure 2.10 for different dimensions of the feature vectors, as this is an important parameter for obtaining better classification performance. The performance on image patches without Gaussian blur is better than that with blur. When $\sigma_{blur} > 3$, the performance drops to approximately 50%. As shown in figure 2.10, even small blur of $\sigma_{blur} = 2$ or 3 affects classification performance.

2.4.3 Results for Synthetic Video Sequences

Hereafter, we evaluate the performance of the proposed smoothing method. Therefore, in this subsection, we assess the performance on synthetic endoscopic video sequences.

We created the synthetic videos as follows. First, we selected three images from the dataset of 1671 NBI images. These were not trimmed patches, but original video frames from which the patches were trimmed. Next, each of these images was repeated 200 times to create a 200-frame static video, resulting in a synthetic video of 800 frames corresponding to four different static scenes. Gaussian blur with $\sigma_{blur} = 5$ was then added into some parts of the videos. Next, we added noise in either of two ways. One was to add Gaussian noise with standard deviation σ_{noise} to every frame, with classification probabilities then obtained using frame-wise classification. The other was to sample Dirichlet noise according to classification probabilities using

$$(2.21) \quad \text{Dir}_{x_t}[\alpha(\hat{\mathbf{y}}_t, s, 1)],$$

where $\hat{\mathbf{y}}_t$ is an observation at individual video frame and s is a scale parameter. The sampled Dirichlet noise was used as an observation vector for each frame.

For training, we randomly selected 300 NBI image patches for each class, 900 NBI image patches in total, from the 1671 NBI image patch dataset. Note that image patches corresponding to images used to create the synthetic video sequences were not used for training.

We should note that the conclusions obtained from the experimental results in this section are limited to observing how fast our method responds to the transitions between blurring and non-blurring frames. This is because the synthetic static videos with blur do not contain any other problematic issues such as abrupt motion or light condition changes. Results for real endoscopic videos are shown in the next section.

2.4.3.1 Comparison of DPF, D-DPF, and Kalman Filter

First, we evaluate the difference between DPF from Section 2.2 and D-DPF from Section 2.3.

Figure 2.11 shows results for a synthetic video to which Gaussian noise has been added. Figure 2.11(a) shows the classification probabilities for each original (noise-free) frame. For this synthetic video, four NBI images were used, each of which lasts 200 frames. We can see three discontinuities at frames 200, 400, and 600. Gaussian blur of $\sigma_{blur} = 5$ is applied to 10 frames before and after the 100, 300, 500, and 700th frame (that is, between frames 90 and 110, and so on) as indicated by shading in figure 2.11(b)–(e). Then, Gaussian noise with $\sigma_{noise} = 1$ was added to all frames to create a final synthetic video for processing. Classification probabilities for this video are shown in figure 2.11(b).

As observed in figure 2.11(b), between frames 200 and 400, the classification probabilities are highly unstable, and for the shaded frames (where blur is applied), the classification probability curves abruptly change. Figure 2.11(c) shows the IPR of each frame, and the values of IPR for the shaded (blurred) frames increase as expected. Results for DPF and D-DPF are shown in figure

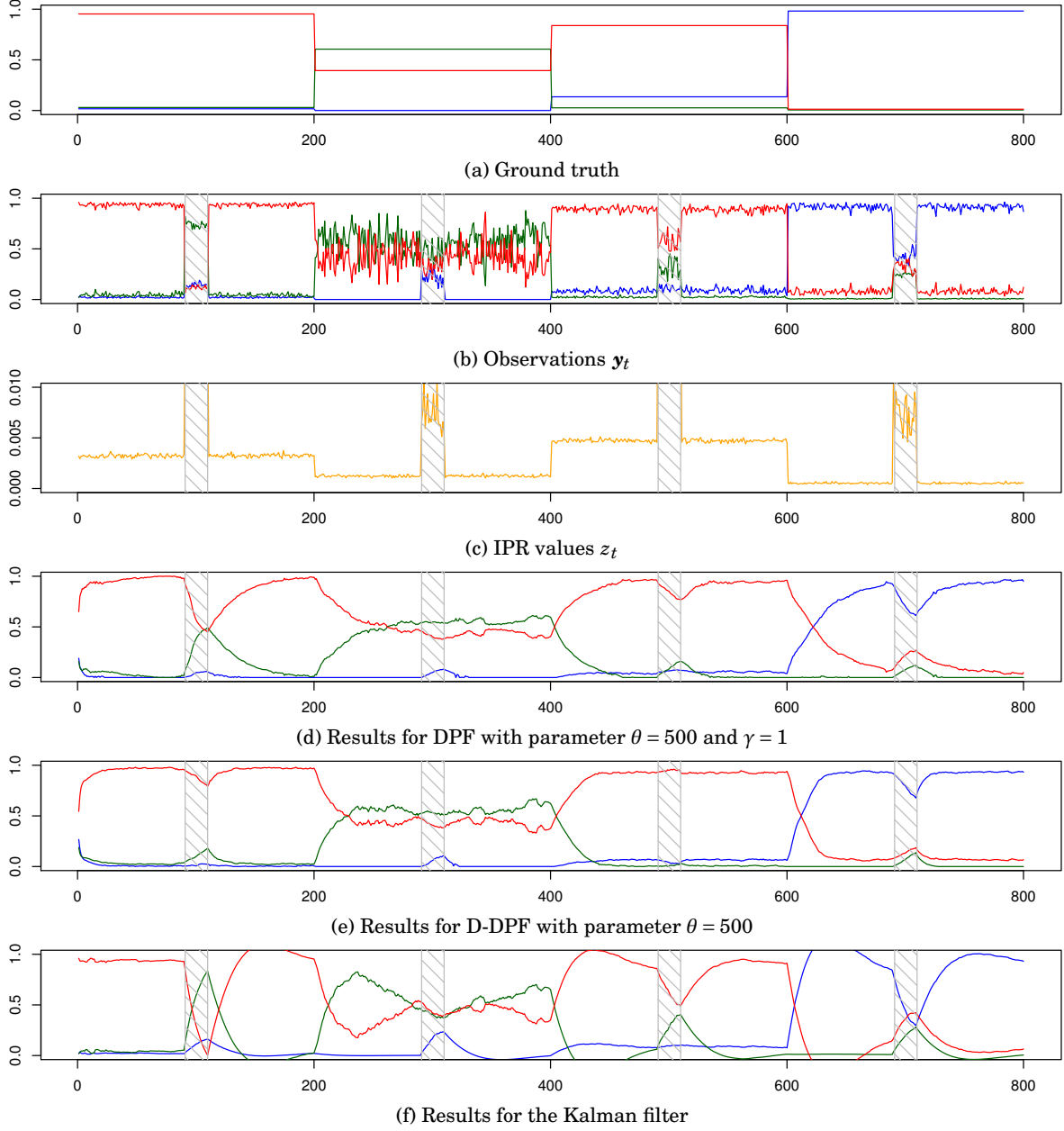


Figure 2.11: Smoothing results on a synthetic video with Gaussian noise of standard deviation $\sigma_{noise} = 1$. The horizontal axis shows frame number. The vertical axis is classification probabilities for the three classes of type A (blue), B (green), and C3 (red) (except (c)). From top to bottom, ground truth of classification probabilities, observations with no smoothing, IPR values, smoothing results for DPF, D-DPF with $K = 1000$ particles, and Kalman filter. Shaded frames are blurred by Gaussian with $\sigma_{blur} = 5$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

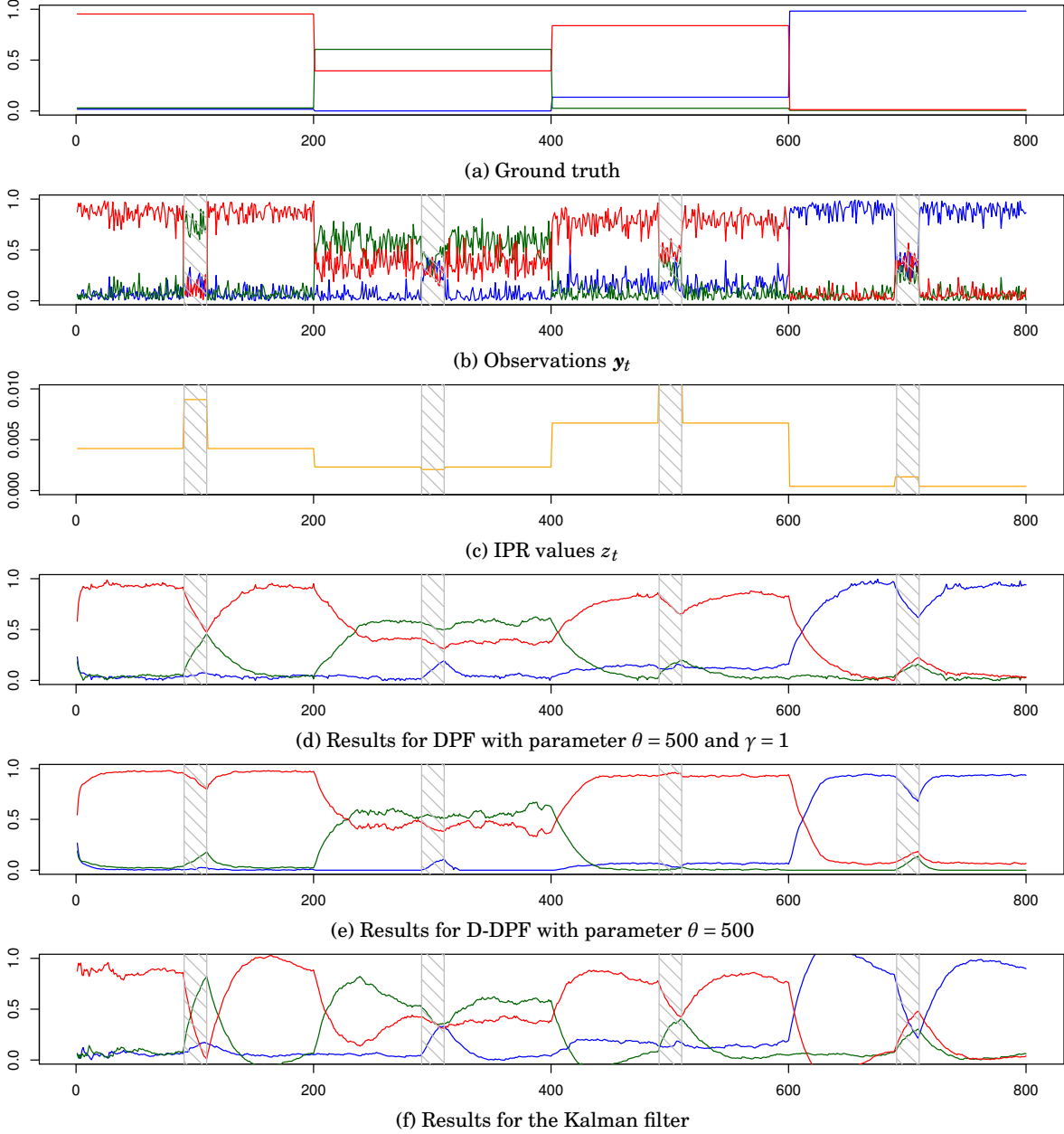


Figure 2.12: Smoothing results for a synthetic video using Dirichlet noise with parameter $s = 20$ (see Eq. (2.21)). The horizontal axis shows the frame number. The vertical axis represents classification probabilities for the three classes of type A (blue), B (green), and C3 (red) (except (c)). From top to bottom, ground truth of classification probabilities, observations with no smoothing, IPR values, smoothing results for DPF, D-DPF with $K = 1000$ particles, and a Kalman filter. Shaded frames are blurred Gaussian with $\sigma_{blur} = 5$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.11(d) and (e), respectively. At approximately frames 100 and 500, where blur is applied, DPF is affected by a sudden change in classification results. In contrast, D-DPF is rather robust to the change due to the defocus information extracted from each frame.

Figure 2.11(f) shows the smoothing result obtained by a Kalman filter. Parameters were manually tuned; hence, the results for the Kalman filter look similar to those for DPF and D-DPF because an optimization with an EM algorithm could not find suitable parameters that would produce satisfactory smoothing results in this case. The Kalman filter was also affected by the sudden change of classification results at frames where blur was applied. Another defect of the Kalman filter was overshooting. Around frames 150, 450, and 650, results exceed the range between 0 and 1. Normalizing or clipping the results in the range of zero to one at each frame would lead to inconsistency with the results for the successive frames, and the probabilistic framework would be lost.

Figure 2.12 shows the results obtained when Dirichlet noise with $s = 20$ has been added to the classification probabilities instead of adding Gaussian noise to the image frames. The procedure for creating the synthetic video was the same. In this experiment, the IPR values computed for the shaded (blurred) frames in figure 2.12(c) are relatively small compared to those for figure 2.12(c). Consequently, D-DPF in figure 2.12(e) is affected much more by the observation. This experiment suggests that a carefully selected model is necessary for the relation between z_t and γ_t .

To obtain a quantitative evaluation, we compute the root-mean-square error (RMSE) between ground truth (figures 2.11(a) and 2.12(a)) and the smoothing results (figures 2.11(d)–(f) and 2.12(d)–(f)) over different amounts of Gaussian or Dirichlet noise. Figure 2.13(a) shows the RMSE for DPF, D-DPF, and the Kalman filter applied to synthetic videos with Gaussian noise for different values of σ_{noise} . Both DPF and D-DPF maintain low values of the RMSE. The RMSE is higher for the Kalman filter than for DPF and D-DPF for an entire range of σ_{noise} values. Figure 2.13(b) shows the RMSE for synthetic videos with Dirichlet noise for different values of s . Note that larger values of s generate smaller amounts of noise. Here again, D-DPF performs better than DPF and the Kalman filter.

2.4.3.2 Results for Different θ

We compare the performance for different values of θ in the state transition. Figure 2.14 shows results for a synthetic video, which was created by the same procedure described above, except that two original images were used to create 200 frames. Then, Dirichlet noise with $s = 50$ was added to the classification probabilities. Figure 2.14(a) shows the classification probabilities for each original (noise-free) frame. For this synthetic video, two NBI images were used, each of which lasts 100 frames. Shading in figure 2.14(b) through (e) indicates frames blurred with Gaussian with $\sigma_{blur} = 5$.

At the discontinuities of frames 50, 100, and 150, smoothed probabilities are pulled to

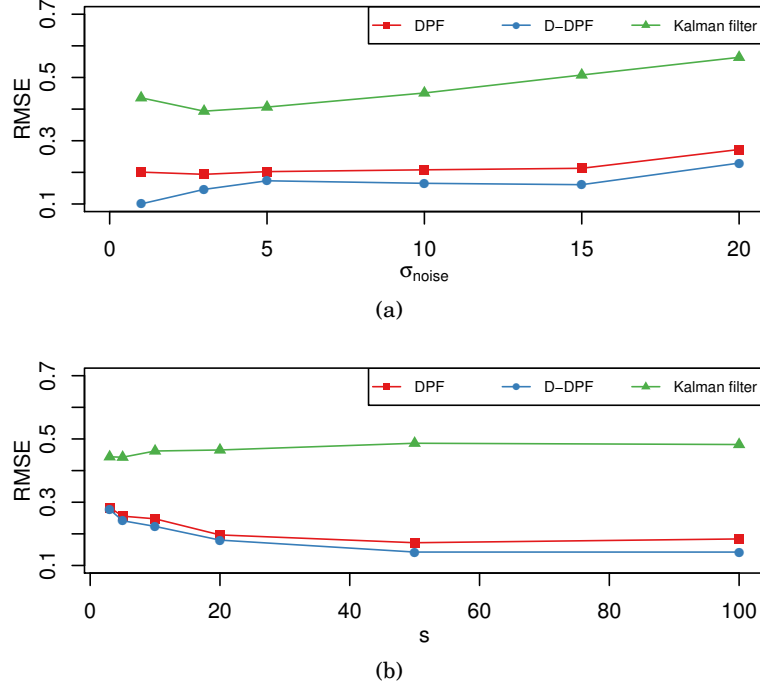


Figure 2.13: RMSE of the smoothing results for the synthetic videos shown in (a) figure 2.11 and (b) figure 2.12. The vertical axis shows RMSE. The horizontal axis is the value of σ_{noise} for the Gaussian noise for (a) and the value of s for the Dirichlet noise for (b) (see Eq. (2.21)).

observations, and θ adjusts the speed of the convergence. When θ is small (e.g., 100), the state transition probability has a broad peak (the middle column of figure 2.3), and successive states $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$ are thus weakly linked; hence, the result rapidly converges to the observation, as in figure 2.14(c). In contrast, when θ is relatively larger (500), the narrow peak of $p(\mathbf{x}_t | \mathbf{x}_{t-1})$ restricts $\mathbf{x}_t$ to be close to $\mathbf{x}_{t-1}$, resulting in a slow convergence such as that in figure 2.14(e).

As a simple extension, one might think of θ as another hidden variable that relates the defocus information and the state transition, as we did with γ for the likelihood. However, we chose not to follow such a direction. If we loosely connected $\mathbf{x}_t$ to $\mathbf{x}_{t-1}$ as well as $\mathbf{y}_t$ when the frame is defocused, then $\mathbf{x}_t$ would not be under the control of either $\mathbf{x}_{t-1}$ or $\mathbf{y}_t$, and the result might be unpredictable. More sophisticated modeling of the relation between the defocus information and the state transition is left as a future work.

2.4.3.3 Number of Particles

We evaluate the results in terms of the number of particles with the same dataset used in the last subsection because the optimal number of particles depends on each problem. Using a large number of particles generally provides good results, but there is a trade off due to increasing computational cost. We fix the parameter $\theta = 100$ and change the number of particles to $K = 10, 100, 1000$, and 10000. As shown in Figure 2.15, the smoothing effect is not sufficient

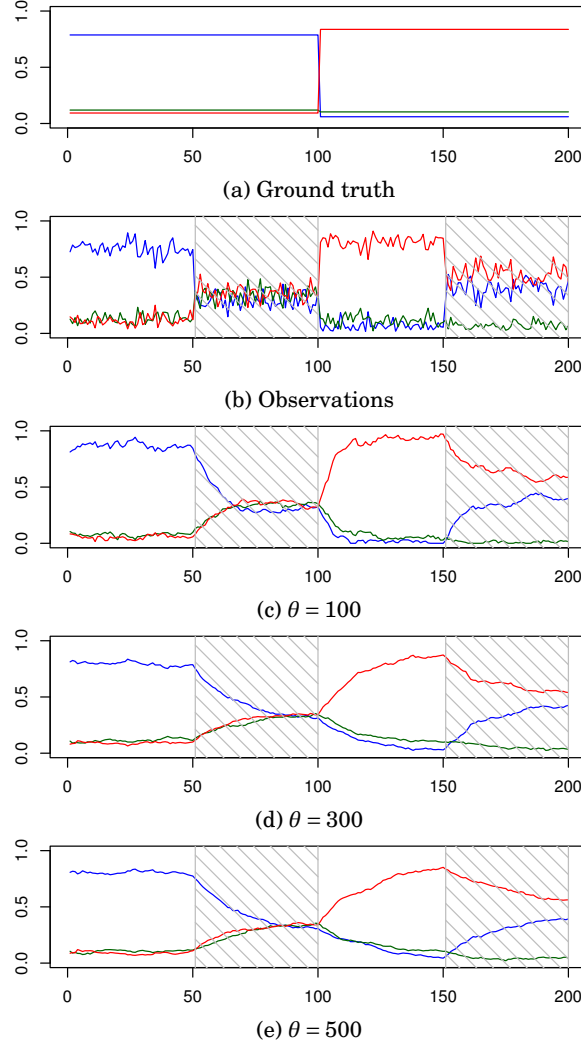


Figure 2.14: Smoothing results for a synthetic video with Dirichlet noise with parameter $s = 50$ (see Eq. (2.21)) with three classes of type A (blue), B (green), and C3 (red). The horizontal axis shows the frame number. The vertical axis shows classification probabilities. From top to bottom: ground truth of classification probabilities, observation with no smoothing, smoothing results with $\theta = 100, 300$, and 500 by D-DPF with $K = 1000$ particles. Shaded frames are blurred by Gaussian with $\sigma_{blur} = 5$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

when using as few as $K = 10$ particles. In contrast, using many particles improves the accuracy of the smoothing results. Evidently, $K = 100$ appears to be sufficient to achieve results comparable to the case wherein many more particles are used, e.g., $K = 1000$ and $10,000$.

Figure 2.16 shows the computational cost per frame, which includes lines 3 to 13 in Algorithm 1. Even when we use $K = 10000$ particles, it takes only 33 ms/frame, of which computing IPR takes 22 ms on average. Furthermore, frame-wise classification requires 50 ms. In total, it requires $88 \text{ ms} \cong 12 \text{ fps}$; this is a sufficient computation speed for a prototype system to be used

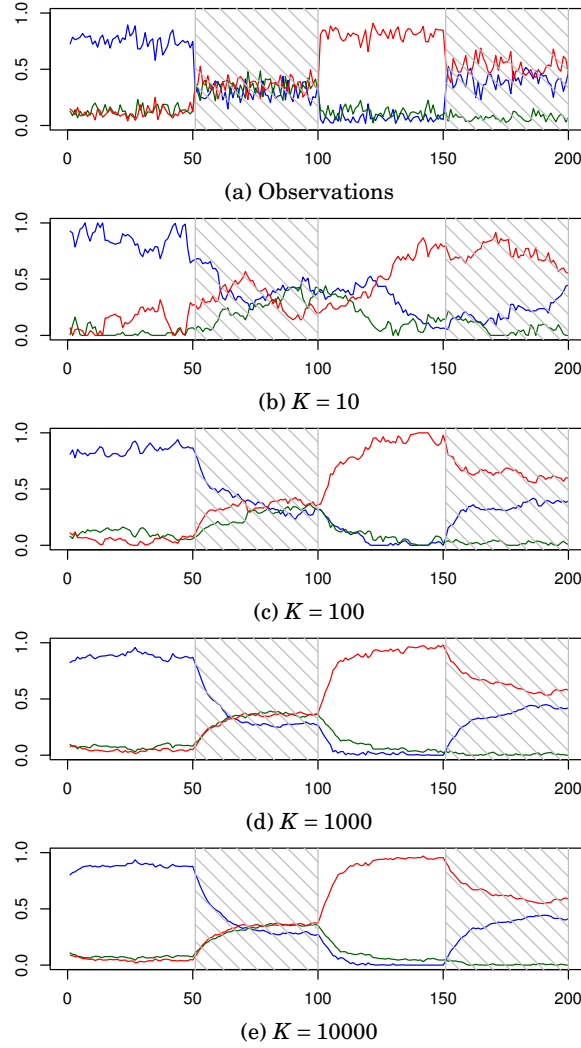


Figure 2.15: Smoothing results for a synthetic video with Dirichlet noise with parameter $s = 50$ (see Eq. (2.21)) with three classes of type A (blue), B (green), and C3 (red). The horizontal axis shows the frame number. The vertical axis shows classification probabilities. From top to bottom: observation with no smoothing, smoothing results with the number of particles $K = 10, 100, 1000$, and $10,000$. Shaded frames are blurred by Gaussian with $\sigma_{blur} = 5$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in diagnosis support during actual endoscopic examinations. Further increase in speed can be achieved by additional fine-tuning of the system. Currently, our unoptimized implementation written in C++ uses a single thread on an Intel Core i5 (2.4 GHz) processor with 16 GB memory.

2.4.4 Results on Real Endoscopic Videos

In this subsection, we demonstrate smoothing results for real endoscopic videos taken during actual endoscopic examinations.

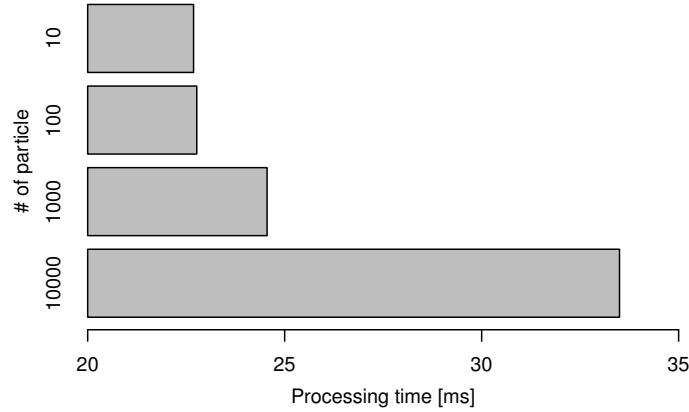


Figure 2.16: Computational cost of D-DPF per frame.

For training, all of the 1,671 NBI image patches in the dataset were used. This dataset is unbalanced, but a preliminary experiment (not shown here) with a balanced dataset of 320 NBI image patches for each class showed results similar to those shown here.

Figure 2.17 shows observation and smoothing results for a video that captures a tumor labeled as type B. During the frames around frame numbers 150, 250, and 450, where observation and IPR values look nearly constant, endoscopists capture the screen to save images of the tumor, and the screen freezes. Due to defocus, type A is dominant between frames 30 and 120, and the observations are unstable particularly around frames 180, 290, 370, and 490. It is evident that the observations (classification probabilities) are highly unstable throughout the video, whereas the results from D-DPF are much smoother. The results obtained from DPF, shown in Figure 2.17(f), are also smooth, but slow to follow the observations. The result from D-DPF with the same parameter shown in Figure 2.17(e) shows a quick follow; particularly, frames between 400 and 500 when observations are close to zero and one. Figure 2.17(g) shows the smoothing result obtained by a Kalman filter with the same parameters as those used for the results in Figures 2.11(f) and 2.12(f). We can see that the results are as slow to follow the observations as DPF. There is also overshooting as we have seen in the last section with the synthetic videos where observations suddenly change such as around frames 150, 250, 450, and 550.

Figures 2.18(a) to (c) show smoothing results for another video labeled as type A, wherein the frames around frame 200 are blurred and the observations are unstable. The results shown in Figure 2.18(c) are smooth and the probabilities for type A have the largest values for all frames as IPR values keep lower values. Figures 2.18(d) to (f) show smoothing results for yet another video labeled as type C3. The results shown in Figure 2.18(f) are smoother than the observations in Figure 2.18(d) as all frames are defocused slightly and IPR values are relatively high.

However, the results in Figure 2.18(f) are type C3 only in frames between 120 and 190 because of the severe instability of the frame-wise classification results. In particular, the results for frames between 60 and 90, and between 190 and 270, are type B because it is dominant

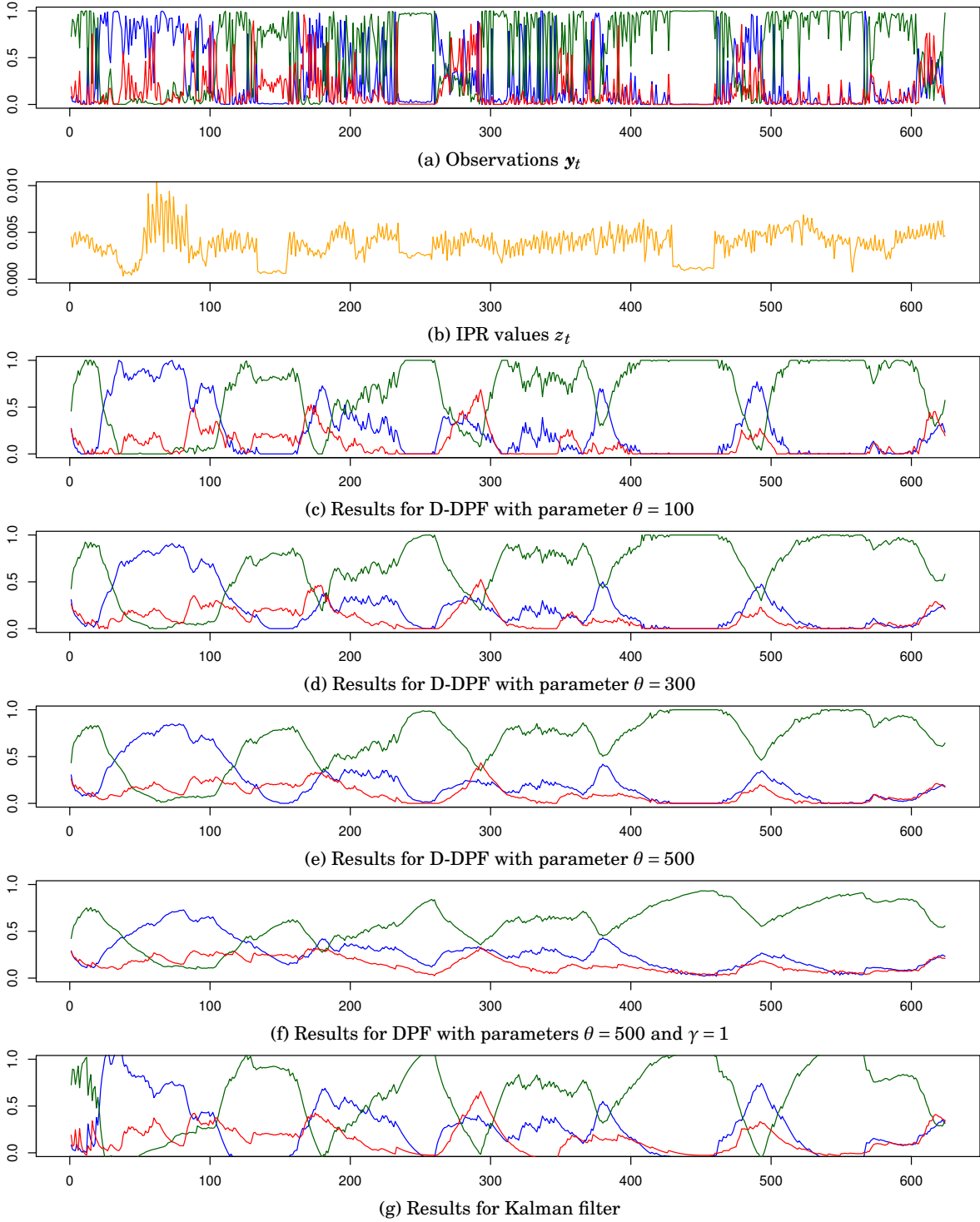


Figure 2.17: Smoothing results on a real endoscopic video of 629 frames labeled as type B. The horizontal axis shows the frame number. The vertical axis shows classification probabilities for the three classes of type A (blue), B (green), and C3 (red) except (b) and the IPR value for (b). From top to bottom, observations with no smoothing, IPR values, smoothing results for D-DPF with parameter $\theta = 100$, 300, and 500, smoothing results for DPF with parameter $\theta = 500$ and $\gamma = 1$, and smoothing results for the Kalman filter.

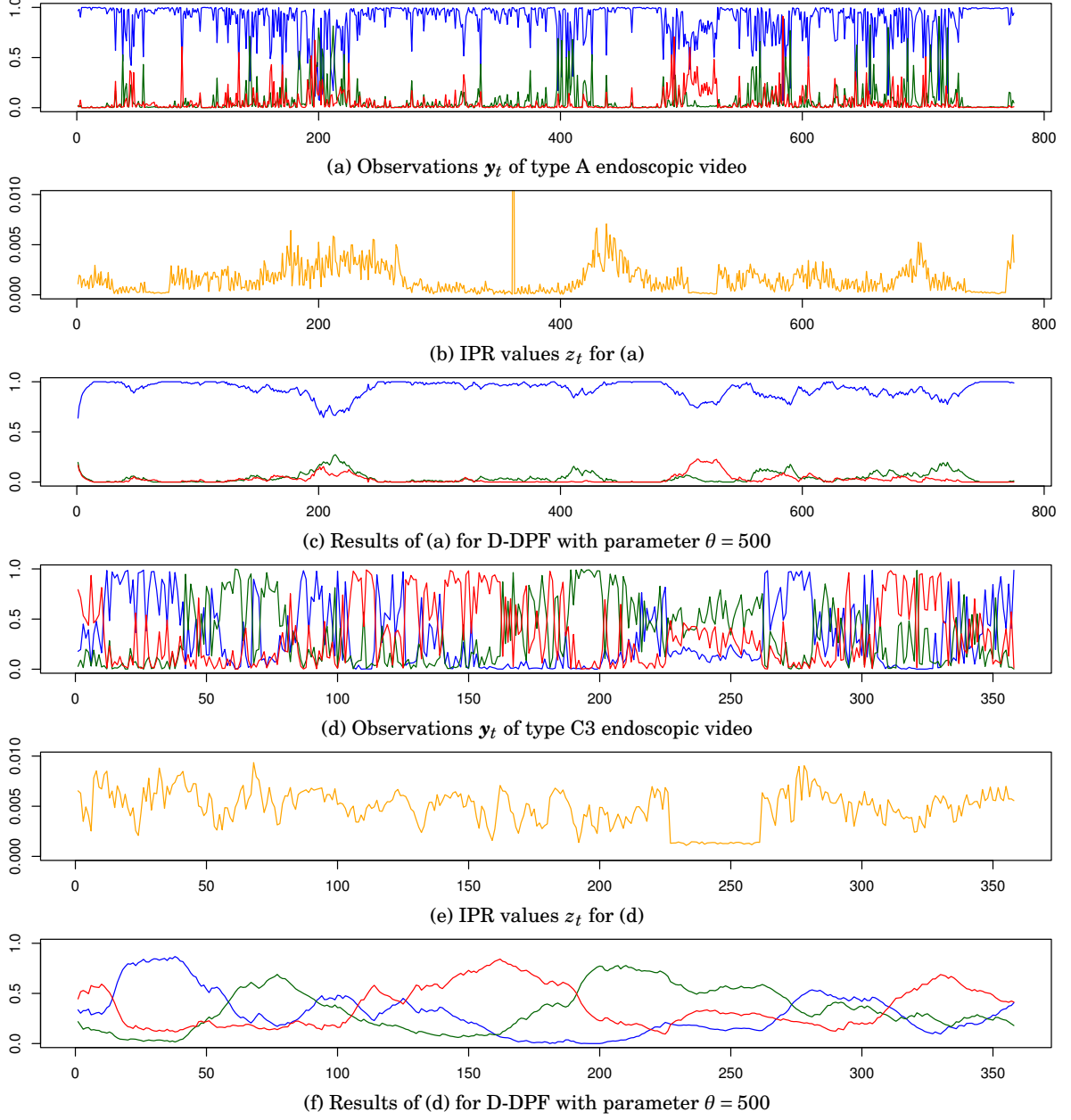


Figure 2.18: Smoothing results on real endoscopic videos labeled as types A (775 frames) and C (358 frames). The horizontal axis shows the frame number. The vertical axis shows classification probabilities for the three classes of type A (blue), B (green), and C3 (red). From top to bottom, observations of a video labeled as type A, IPR values for (a), smoothing results for (a) by D-DPF, observations of a video labeled as type C3, IPR values for (d) and smoothing results for (d) by D-DPF. Parameter θ of D-DPF is set to 500.

in the observations during these frames. This may be caused by defocus of frames, as well as other problematic issues that occur in real endoscopic videos such as illumination change, color bleeding, and abrupt camera motion.

2.5 Summary

We have proposed a novel method –D-DPF– to smooth the classification probabilities obtained from frame-wise endoscopic image classification by incorporating defocus information into a particle filter with a Dirichlet distribution. We assumed that the defocus information extracted from each frame influences classification probabilities, and we proposed linking the Dirichlet likelihood to the defocus information and the IPR proposed by Oh et al. [120], which is a ratio of the number of edge pixels isolated from neighbor edge pixels. Then we sampled parameter γ_t in the likelihood from a Rayleigh distribution.

For endoscopists, unstable recognition results such as those in Figure 1.11 are difficult to use for diagnosis. The proposed smoothing method improves the visibility and understandability of the recognition results and facilitates the use of the results for diagnosis. Moreover, the proposed method has the potential to be used for training endoscopists who have less experience of endoscopic examinations.

D-DPF can be extended in several ways. One possible extension is to address other causes of instability. We have focused on defocus information herein, but other causes also exist. One example is color bleeding due to the following property of NBI endoscopes: different wavelengths of light are used to create a single frame by rotating a filter in front of the light sources. Thus, color bleeding (i.e., different color illuminations appear at the same time) occurs when the assumption that the scene is temporally static is violated owing to the large motion of the endoscope. Rapid movement of the endoscope also results in motion blur, another cause of instability. The proposed D-DPF might still be applicable in such situations if we could introduce metrics representing color bleeding or motion blur instead of defocus information. For other frame-wise classification results with four or five-classes, we can apply D-DPF by simply changing the dimension N of the Dirichlet distribution. Gastrointestinal endoscopic videos are also in the application range of D-DPF. Given additional information along with the signals to be smoothed, effective smoothing results can be obtained.

How to visualize the results more effectively is another issue that deserves further attention. For every frame, we compute the mode of the Dirichlet distribution estimated in the update step as the smoothed classification probabilities. Furthermore, the estimated label (the class having the largest probability from the estimated mode) is displayed as a colored rectangle shown at the patch used by the frame-wise recognition. Other possible means of visualization include displaying classification probability curves that are similar to an electrocardiogram or visualizing the estimated Dirichlet distribution shapes instead of probabilities and labels. In any

case, further consideration is needed in terms of human-computer interaction.

In addition to the visualization issue, our future work includes embedding the proposed method into an actual working system for clinical evaluations. We also must explore alternative ways to represent defocus information (other than IPR) and other sampling strategies (apart from the Rayleigh distribution) for the likelihood parameter.

SVM-MRF IMAGE LABELING OF NBI ENDOSCOPIC IMAGES

In this chapter, we extend our previous work [145, 146] for NBI image classification to image labeling (or segmentation): trying to find which part of the NBI image falls into one of three types of colorectal tumors. The previously developed system was based on a BoVW framework with densely sampled SIFT descriptors (see Figure 1.7). Each training image was transformed into a histogram of visual words, then classified by an SVM classifier with a linear kernel. To build an image labeling method on top of our previous classification system, we combine SVM classifiers with a MRF minimization framework; first we divide an NBI image into a number of small square patches, then classify each patch by SVM classifiers that are separately trained, and finally make the classification results for the patches spatially consistent by minimizing an MRF energy function. In experimental results, labeling results obtained by the developed system are evaluated on a dataset of colorectal NBI endoscopic images.

3.1 Related Work

The combined use of SVM and MRF has been already explored in the literature and some relevant work is reviewed here.

Wu et al. [157] proposed a prior feature SVM-MRF, a labeling method for 3D Magnetic Resonance images. They trained SVMs using pixel locations in addition to the voxel intensity and the given label of the point, then used the posteriors (outputs) of the SVMs for the pixels in the unary term of the MRF energy, which was minimized by the Iterated Conditional Modes (ICM). Wang and Manjunath [156] proposed an SVM-MRF framework for image retrieval based on semantic segmentation of images. For each block, SVMs output conditional probability which is used as the unary term of the MRF energy. They used a causal greedy algorithm to minimize

the energy. Moser and Serpico [110, 111] used binary SVMs with a particular kernel to define a local decision rule of MRF with updates by ICM. Hoefel and Elkan [64] proposed a two-stage SVM/CRF for sequence (hence chain) classification. Discretized scores of SVMs, which are trained separately, are used as a term in the CRF energy function, and the weights of the terms are learnt.

Most relevant to ours is the work of Fulkerson et al. [33]. They trained SVMs by bag-of-feature histograms, and then scores by SVMs for each pixel are used as confidence values for localizing objects. Later, they extended the approach by incorporating spatial context with conditional random field (CRF) [34]. SVMs are trained with bag-of-feature histograms of super pixels, and then probability outputs by SVMs for each superpixel are used as a unary term of CRF energy. To handle boundaries of objects, neighboring histograms of a superpixel are merged, then more accurate object boundaries are inferred by the CRF energy minimization with α -expansion. The pairwise term uses color difference and edge length between superpixels. The parameter learnt as a CRF model is a single weight between unary and pairwise terms.

Our objective is to segment an NBI image into three types (A, B and C3) of the NBI magnification findings (Figure 1.6). This labeling problem is very different from the targets of the existing works mentioned above, or other works on colon polyp segmentation [4, 5, 13, 41]. First, we can not use the location and rotation of colorectal tumors in NBI images because those are taken by an endoscopy that arbitrarily moves around in the colon. Therefore, a location prior [157] in an image is not useful. Second, there is no clear borderline between the types; instead, visual appearance of texture of the mucosal surface changes gradually from one type to another. Edge information is usually useful for labeling of polyps [4, 5, 13, 41], however we do not expect so in our task. We aim to assign labels conditions of cancer to pixel. To avoid confusion, we mention the term *segmentation* for finding contours and *labeling* for assigning pixel labels like us our task. In the next section, we introduce an SVM-MRF combination similar to [34], but without edge information and superpixels.

3.2 SVM-MRF Image Labeling

Here we describe an image labeling method using SVM and MRF. First, we divide an NBI image into square patches P_i ($i = 1, \dots, n$), and each patch corresponds to a site (or node) of a grid of MRF. Adjacent patches may overlap depending on the size of patches and the spacing of the MRF grid.

Each site i has a label x_i taking values of A, B, or C3 corresponding to the three types of the NBI magnification findings. We denote all the labels collectively as $\mathbf{x} = (x_1, \dots, x_n)$. The bag-of-visual word histogram of the patch is denoted as y_i , and collectively $\mathbf{y} = (y_1, \dots, y_n)$.

We define the following MRF energy function in terms of the posterior probability:

$$(3.1) \quad f(\mathbf{x}|\mathbf{y}) \propto \exp \left(\sum_i A(x_i, y_i) + \sum_{j \in N_i} I(x_i, x_j) \right).$$

Here $A(x_i, y_i)$ is a unary term that represents the inconsistency of the patch label x_i to data y_i . In this study, we use the posterior probability outputs $\log P(x_i|y_i)$ of SVM classifiers that are learnt with a separate training set of NBI images. Then we use the posterior as the unary term as follows:

$$(3.2) \quad A(x_i, y_i) = -\log P(x_i|y_i).$$

The second term $I(x_i, x_j)$ in the MRF energy is called an interaction term which describes the spatial inconsistency between x_i and its neighbors x_j , where N_i is a neighbor of site i . Here we define the interaction term as follows:

$$(3.3) \quad I(x_i, x_j) = \begin{cases} -\log p, & x_i = x_j \\ -\log \frac{1-p}{2}, & \text{otherwise} \end{cases},$$

where $p \in (0, 1)$ is a probability that site i and its neighbor j take the same label. Because the labeling problem here has three labels (A, B, and C3), there are two cases where site i and its neighbor j take different labels. Therefore the probability $1 - p$ is halved.

The MRF energy function (3.1) is minimized by α - β swap Graph Cuts [11] in order to obtain an MAP estimate of labels $\mathbf{x}$.

3.3 Considering Highlight Regions

Using the aforementioned MRF model, we can segment endoscopic images into three classes smoothly. However, there exists a lot of unnecessary regions for diagnosis such as highlights and blurred background. To suppress these regions, we introduce a model considering highlight regions with wrong local features around highlight regions.

3.3.1 Detecting Highlight Regions

To detect highlight regions, we use the method proposed by Oh et al. [120]. They detect highlight regions in two ways: first one is to use simple thresholding for HSV color space. Second one is to divide images into several regions by JSEG [24], and then, detect relative highlight regions for each region using box-plot of pixel values and summary statistics such as upper quantile, lower quantile, minimum and median values.

In this research, we use only the first method, that is, simple thresholding method. At first, we convert endoscopic images from RGB to HSV color space where $H \in [0, 359]$, $S \in [0, 1]$, $V \in [0, 1]$. Then, applying for saturation (S) and Value (V), we can detect highlight regions. If S is lower than 0.35 and V is higher than 0.75, these pixels are highlight regions. Figure 3.1 shows examples of highlight regions.

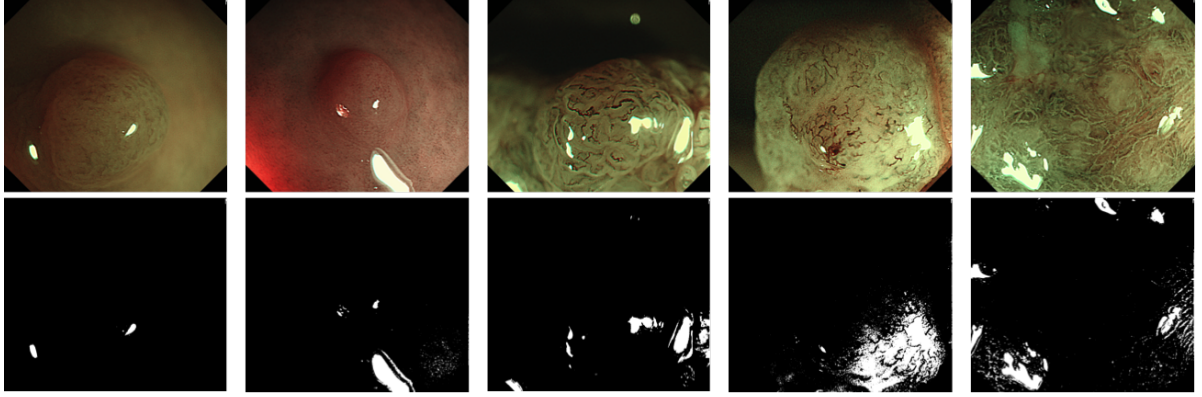


Figure 3.1: Results of detecting highlights. Upper row: original images. Lower row: detected highlight.

3.3.2 MRF Model Considering Highlight Regions

We define the following MRF energy function with highlight information:

$$(3.4) \quad f(\mathbf{x}|\mathbf{y}, \tilde{\mathbf{y}}) \propto \exp \left(\sum_i A(x_i, y_i, \tilde{y}_i) + \sum_{j \in N_i} I(x_i, x_j) \right),$$

where $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_n)$ is highlight information of each patch which takes 1 or 0 if the patch has highlight region or not.

Unary term $A(x_i, y_i, \tilde{y}_i)$ can be formulated as follows:

$$(3.5) \quad A(x_i, y_i, \tilde{y}_i) = \begin{cases} -\log P(x_i|y_i) & \tilde{y}_i = 0 \\ -\log a & \tilde{y}_i = 1 \end{cases}.$$

We don't trust the visual-word histogram y_i of a patch in highlight regions and ignore the posterior probability outputs of SVM classifiers. In other words, we take the same energy values a for those to ignore labels. We use the same binary term $I(x_i, x_j)$ with energy in Section 3.2.

3.4 Experimental Results

3.4.1 Dataset and Evaluation Method

We have 1671 NBI images (Type A: 504, Type B: 847, and Type C3: 320), which were collected during NBI colonoscopy examination. Each image was trimmed from an original NBI image to a rectangle as a training image representing typical microvessel structure appearance, and was labeled by medical doctors and endoscopists. We selected 160 images from each type for a training set; totally 480 training images are used for training an SVM classifier. For test, the remaining 1191 NBI images (Type A: 344, Type B: 687, Type C3: 160) were used. In each NBI image, an

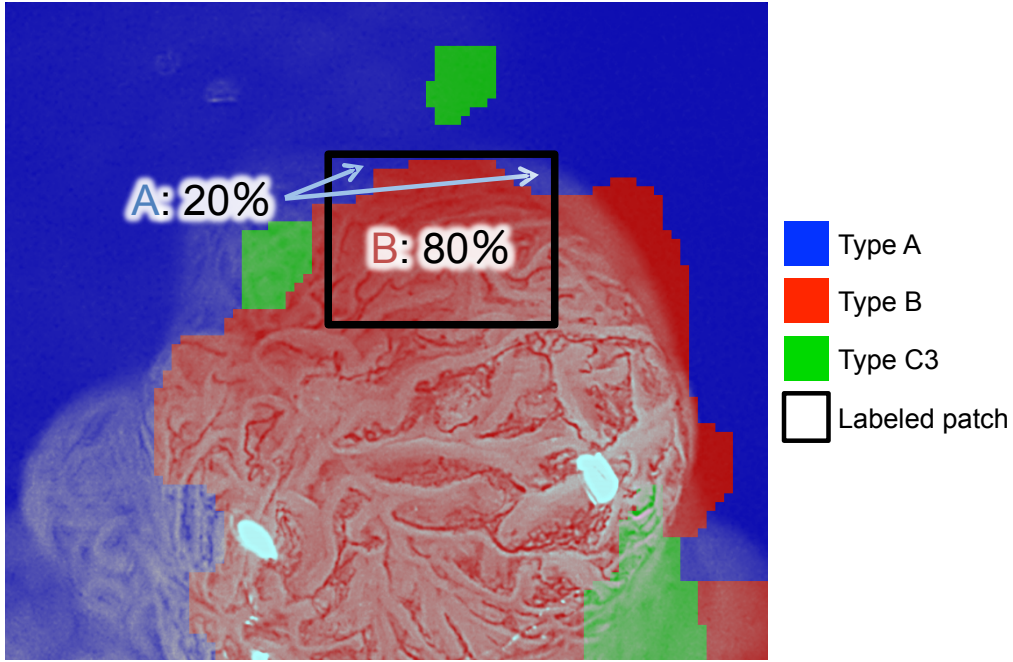


Figure 3.2: Evaluation procedure (best viewed in color). Performance is evaluated by the ratio of areas inside a rectangle whose estimated labels are correct. In this example, the ground truth of the rectangle is Type B and 80% of the area inside the rectangle is correctly labeled.

MRF grid of spacing 10 pixels is constructed. A node of MRF corresponds to a square patch of the size of 120 pixels, in which SIFT descriptors are extracted at each 5 pixels with fixed scales of 5 and 7 pixels [145].

Image labeling methods are usually evaluated by ground truth labels of an entire image. This means that doctors need to paint a lot of NBI images but this is impractical. Instead, we use the above-mentioned trimmed regions from the original NBI images for training the SVM classifiers, because the labeled rectangle in the original NBI image is a good indicator how good the labeling result is. As shown in Figure 3.2, we evaluate a labeling result by the ratio of areas inside a labeled rectangle whose estimated labels are correct. Using the estimated labels in the labeled region, correct rate, precision rate and recall rate are calculated from a confusion matrix [127].

3.4.2 Labeling Results without Highlight Regions

Figure 3.3 shows performances of labeling results over the probability p values of 0, 0.05, ..., 0.95, 0.99. We can see that the correct rate improves as p becomes large; the smoother the labels of the adjacent patches, the better the result.

Figures 3.4 to 3.6 show examples of labeling results for each NBI type. In all cases, resulting regions become smoother and large as p becomes large, and particularly a good labeling result is obtained in Figure 3.5.

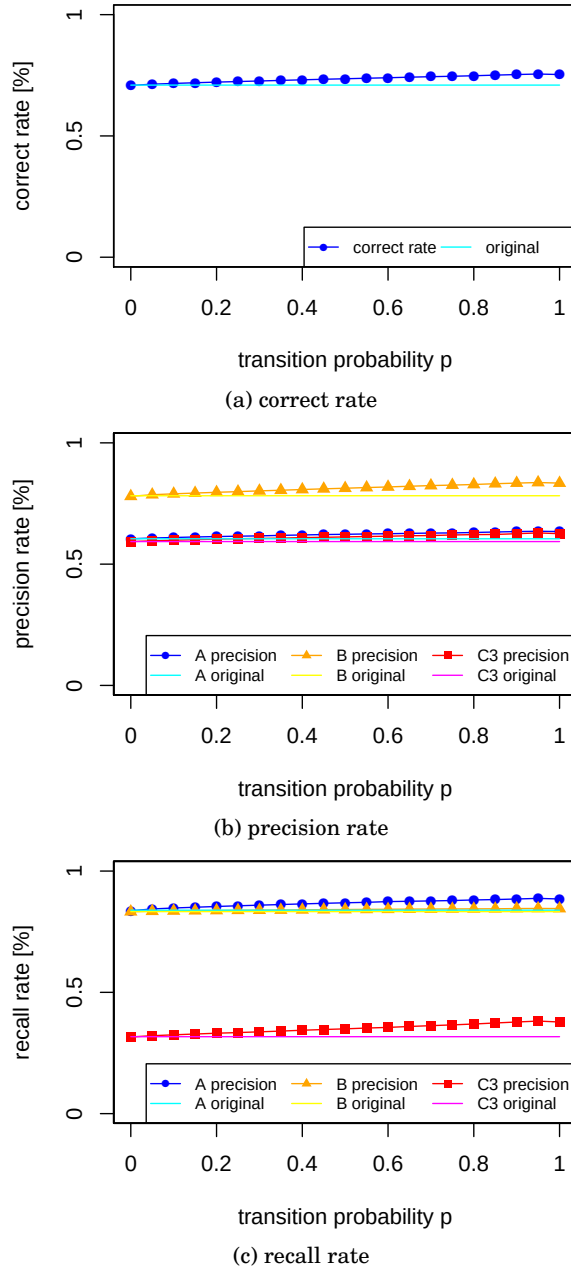


Figure 3.3: Performance of labeling results for different values of p (best viewed in color) in terms of correct, precision and recall rates for each type. "Original" means results obtained when MRF is not used and each patch is independently classified by an SVM.

Note that the part around highlights due to the reflection of light in Figure 3.4 is classified as Type B while the true label is Type A. This may be caused by the strong edge of the highlight; many edges can be seen in Type B images while textures in Type A images are rather smooth. These effect will be investigated in the next experiment.

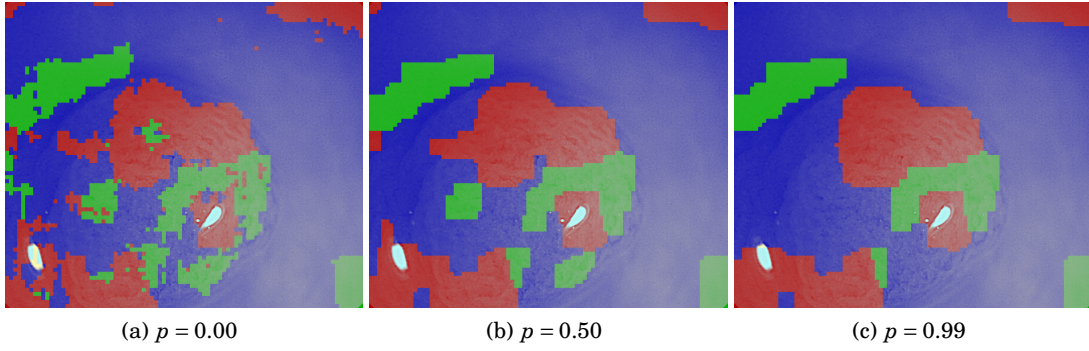


Figure 3.4: Labeling result for NBI images of Type A for different values of p (best viewed in color). Blue color represents Type A, red Type B, and green Type C3.

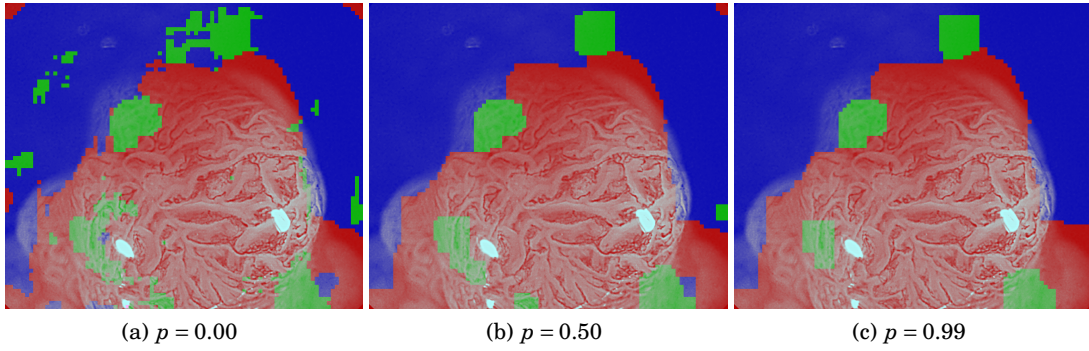


Figure 3.5: Labeling result for NBI images of Type B for different values of p (best viewed in color).

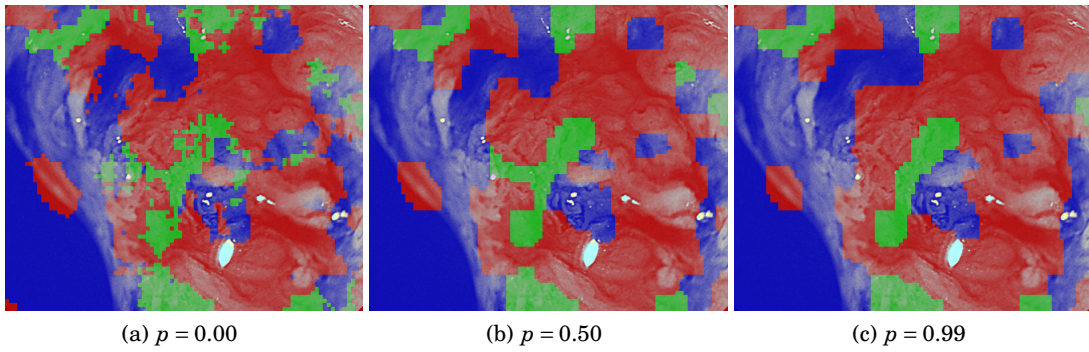


Figure 3.6: Labeling result for NBI images of Type C3 for different values of p (best viewed in color).

3.4.3 Labeling Results with Highlight Regions

Figure 3.7 to 3.13 show labeling results with highlight regions. Using highlight regions, we found that the holes where the highlight exists are removed. Type A and C3 images tend to be classified into type B at the highlight regions. In contrast, type B images tend to be classified into type A.

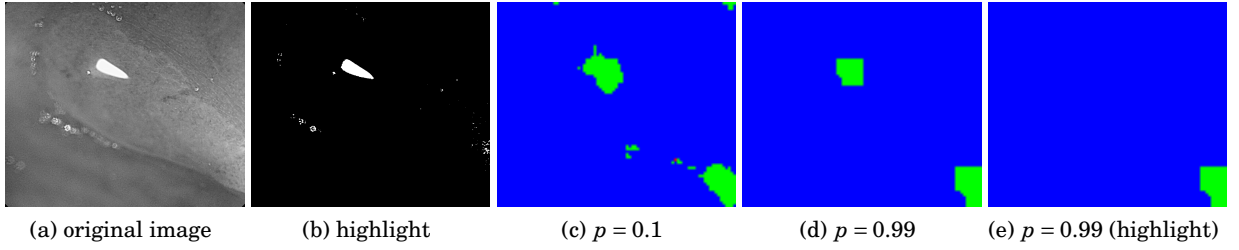


Figure 3.7: Labeling result for NBI images of Type A-1 for highlight regions (best viewed in color).

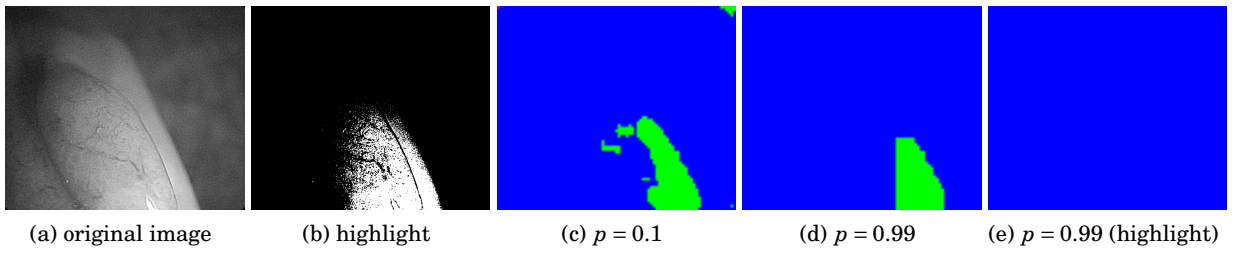


Figure 3.8: Labeling result for NBI images of Type A-2 for highlight regions (best viewed in color).

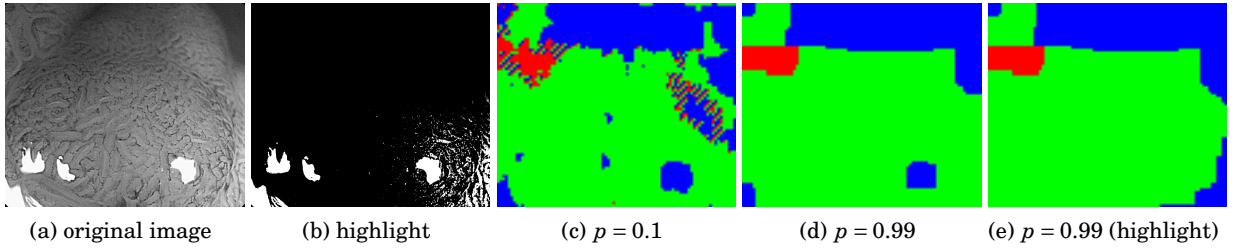


Figure 3.9: Labeling result for NBI images of Type B-1 for highlight regions (best viewed in color).

3.5 Summary

In this chapter, we presented an image labeling method based on a SVM-MRF combination. Currently the parameter p , the probability that two adjacent patches takes the same label, is left for operators to tune how much the labeling result is sensitive to noise. Then, we introduce the highlight information into our MRF energy minimization framework. Experimental results demonstrate the effectiveness of the proposed method and the influence of noise. Future work includes adding noise information such as out of focus, automatic adjustment of the parameters, and labeling of NBI endoscopic video sequences.

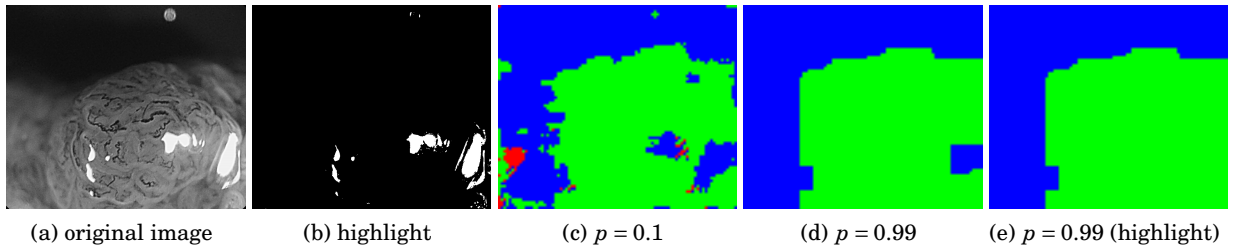


Figure 3.10: Labeling result for NBI images of Type B-2 for highlight regions (best viewed in color).

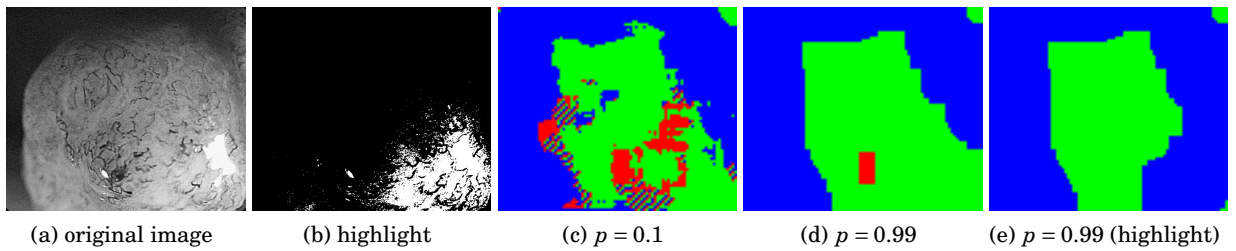


Figure 3.11: Labeling result for NBI images of Type C3-1 for highlight regions (best viewed in color).

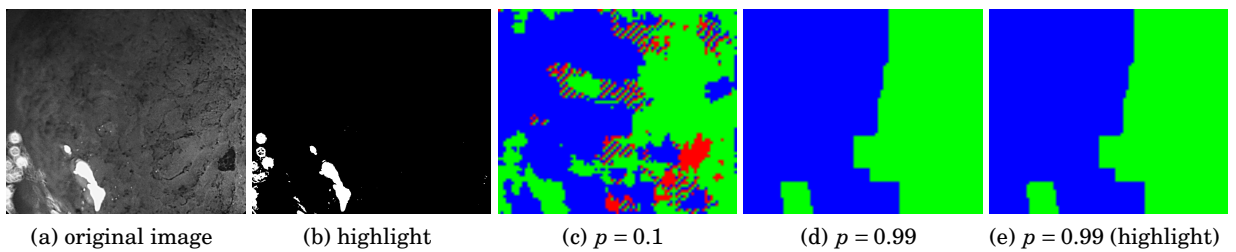


Figure 3.12: Labeling result for NBI images of Type C3-2 for highlight regions (best viewed in color).

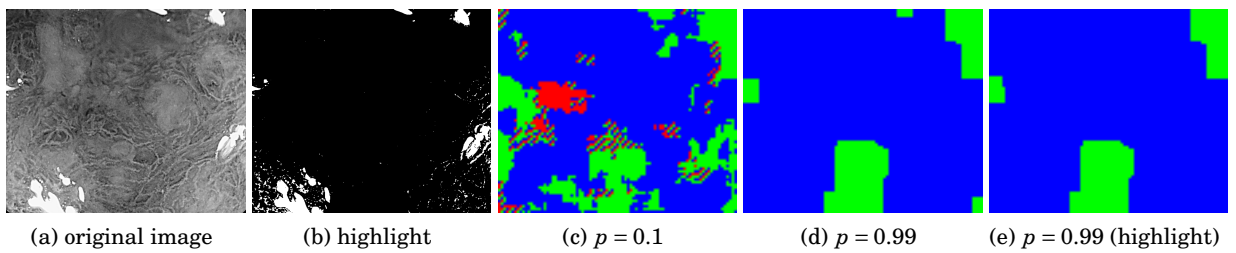


Figure 3.13: Labeling result for NBI images of Type C3-3 for highlight regions (best viewed in color).

TREE-WISE DISCRIMINATIVE SUBTREE SELECTION

This chapter investigate another approach for image labeling task. Also for the same purpose, we proposed an image labeling method based on SVM and MRF in Chapter 3, but the obtained results were not satisfactory enough. One reason is the lack of spatial consistency of MRF framework. In general, object shapes and boundaries are roughly modeled by the pairwise term of an MRF model with edges in the image. However, NBI endoscopic images used in our task often do not have clear boundaries between categories and therefore it would be difficult to model the edge information by the MRF. Another reason lies in the large variation of the texture caused by geometrical and illumination changes. Colorectal polyps and intestinal walls are not flat but undulating (wave-like or spherical shapes). Furthermore, endoscopic images have high contrast textures due to the lighting condition of the endoscope. In such a circumstance, recognition methods would fail because texture descriptors such as visual word histogram. Nevertheless, texture is one of the most important cues for image understanding. A number of texture descriptors, such as wavelet transforms [101, 102], local binary patterns (LBPs) [121], Gabor [10, 152], and textons [89], have been proposed to model texture in images, and image labeling and segmentation methods using such texture descriptors have been proposed. One of the limitations of these texture descriptors is the difficulty of representing a wide variety of changes in texture appearance. More specifically, texture appearance changes in geometry, scale, and contrast. Therefore, learning-based image segmentation methods must have a large number of training images with ground truth segmentation labels to be able to adapt to the texture variations. However, these methods are not applicable to cases with a small number of training images.

In this chapter, we propose a method for texture image segmentation that works with a small number of training images. In some cases, such as segmentation of natural images, we might

be able to collect and use a large number of training images, whereas in other cases, such as in medical image analysis, collecting data and creating ground truth labels are very expensive and difficult.

There has been a promising attempt to deal with geometrical and contrast changes in texture. Xia et al. [158] have proposed a texture descriptor, shape-based invariant texture analysis (SITA), for a texture image classification task based on the tree of shapes [35, 109]. In the field of mathematical morphology, a hierarchical representation, the morphological tree, is popular, and a number of hierarchical trees, such as min/max trees [69, 112], binary partition trees [131], minimum spanning forests [20], the tree of shapes [35, 109], and color tree of shapes [17], have been proposed. Morphological trees have been applied to, for example, biomedical imaging [26, 126, 159]. Xia et al. [158] focused on the natural scale-space structure and invariance of contrast change in the tree of shapes and proposed the SITA descriptor based on the tree of shapes (details are described in Section 4.2.2). To the best of our knowledge, this was the first attempt to create texture descriptors from the tree of shapes. A SITA feature is a histogram of texture features aggregated from all of the nodes in a tree, with the root node of the tree representing the SITA feature of the image. It can be noted that a node corresponds to a part (or a region or blob) of the image, whereas the root represents the entire image. A parent node corresponds to a blob that contains blobs of children nodes. This constitutes a hierarchical structure of the image, which is called the tree of shapes. Xia et al. [158] show through their experimental results that this hierarchical structure renders SITA features invariant to local geometric, scale, and radiometric changes, with good performance in image classification and retrieval problems. Other texture descriptors based on the tree of shapes have also been proposed. Liu et al. [94] introduced a bag-of-words model of the branches in a tree of shapes and represented the co-occurrence patterns of shapes. He et al. [55] adopted the basic idea of LBPs to propose a texture descriptor. However, these methods handle only texture patch classification and retrieval tasks, and no work has been performed on handling multiple textures in a single image for texture segmentation.

Inspired by the invariance property of SITA, in this study, we propose a novel segmentation method for texture images. An overview of the proposed method is shown in Figure 4.1. The idea of our method is to adopt SITA, but to use it for segmentation of an image rather than for classification of images. In the original work on SITA [158], a SITA feature is computed for a classification task at the root of the tree of an image. Here, for a segmentation task, we compute the SITA features at all of the nodes in the tree and classify every node to predict labels of pixels corresponding to the nodes. In other words, we compute SITA features at root nodes of all of the subtrees of the original tree. This simple concept is rather straightforward but has a problem of instability for histogram feature computation. If we compute a SITA feature at the root of a small subtree, for example, near the leaf nodes of the original tree, then the resulting histogram (i.e., the SITA feature) is less stable and discriminative for classification because the small subtree has a small number of nodes available for SITA histogram computation. For this reason, we

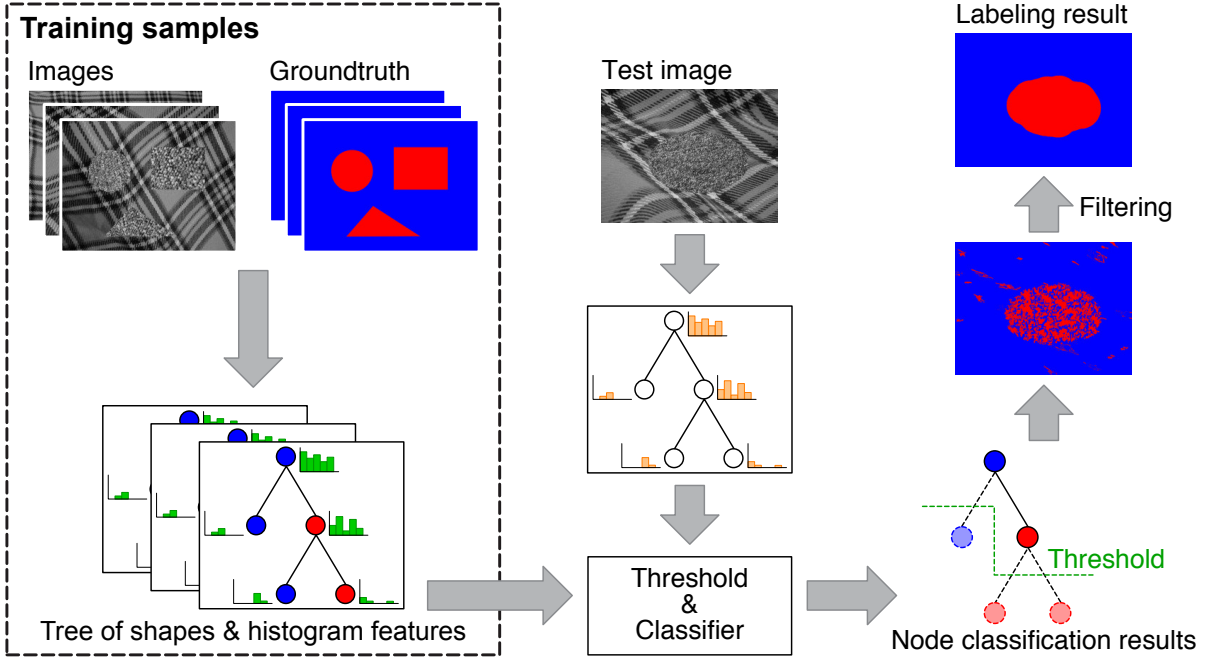


Figure 4.1: Overview of the proposed method.

propose to find subtrees that are sufficiently stable and discriminative for classification by jointly estimating the sizes of subtrees and training a classifier in the training stage. At the labeling stage, given a test image, we estimate (again simultaneously) the sizes and labels of the subtrees of the tree of the test image.

The contribution of our work is two-fold. First, we propose a novel image segmentation framework based on the tree of shapes that makes our method robust to changes in texture appearance. Second, our method works on a small dataset of training images. We use the SITA features of many nodes from an image instead of a single SITA feature at the root node per image. Therefore, our method learns sufficient features to be discriminative for training, whereas the number of training images can still be small (details are described in Section 4.4).

4.1 Related Work

Related works on image labeling and segmentation of colorectal images are already introduced in Section 3.1. In this section, we briefly introduce major image labeling frameworks.

Texture image labeling (or segmentation) is a well-studied task in the field of computer vision, medical imaging [58, 148, 162], and synthetic aperture radar (SAR) image processing [23, 155], and a number of methods for performing that task have been proposed. One popular approach uses MRFs. For modeling spatial consistency, an MRF comprises unary data terms of individual pixels (patches, sometimes super pixels) and pairwise terms between neighbors. The accuracy of MRFs highly depends on the unary term, for which various texture descriptors are used. A

number of unsupervised texture image segmentation methods based on MRF have been proposed [74, 149, 150], but we focus here on supervised texture image segmentation methods using MRFs. One of the supervised MRF approaches was introduced in chapter 3. They proposed a patch-based method that uses a posterior probability obtained from an SVM with BoVW histograms using more than 1,000 labeled patches for training.

Another popular approach uses CRFs; MRF is a generative model, whereas CRF is a discriminative model. CRF has a structural-learning property and can train the spatial structures of labels. Shotton et al. [137, 138] proposed TextonBoost for object segmentation. They introduced a novel texture descriptor, the texture-layout filter, into the CRF framework. For evaluation, they used the MSRC-21 dataset and 271 images [138] for training. Bertelli et al. [6] adopted a kernel-structured SVM for object segmentation. They introduced a pairwise term to a structured SVM that is the same as the CRF framework. Their method was evaluated via 3-fold cross validation with three datasets [9, 115]. For each trial, 400, 218, and 56 images for the three datasets, respectively, are used for training. A fully connected CRF has been proposed [80] that uses a mean field approximation with a linear combination of Gaussian kernels for a pairwise edge potential for efficient inference. In their experiment, they used approximately 270 images on the MSRC-21 [138] dataset and 770 images on the PASCAL Visual Object Classes (PASCAL VOC) dataset [28] for training.

Recently, convolutional neural networks (CNNs) have been proposed for computer vision tasks that include image labeling. Farabet et al. [30] proposed a labeling method for scene parsing. Their approach assigns estimated labels to pixels and then refines the results using superpixels, CRFs, and optimal-purity covers on a segmentation tree. Long et al. [97] used a fully convolutional network trained in an end-to-end manner. These two methods use the SIFT flow dataset [92] for evaluation, with 2,488 images used for training and 280 images used for testing. Other methods have also been proposed [36, 93, 117] using hundreds of images for training. Consequently, these CNN-based approaches have shown good performance, as long as large numbers of training images are available. It would be difficult to achieve good performance for smaller datasets.

In contrast to the methods above, our proposed method works effectively with a small number of training images. In this study, we show a comparison of the proposed method with these related approaches using a small dataset of texture image labeling.

4.2 Tree of Shapes and SITA

Herein, we briefly describe the definition of tree of shapes and the SITA histogram feature.

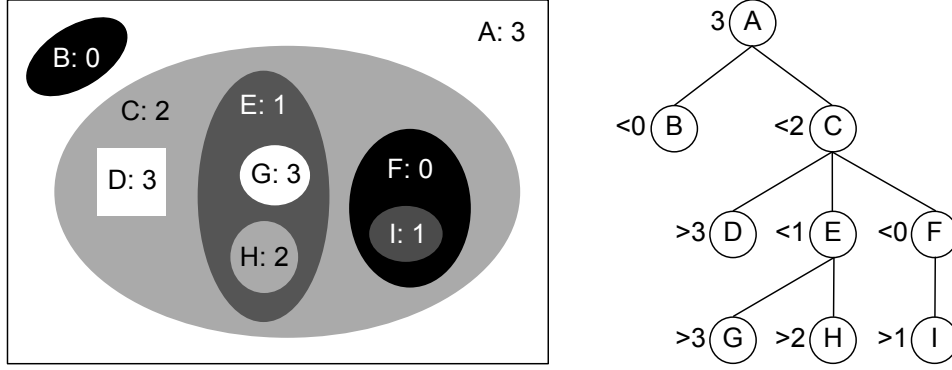


Figure 4.2: Example of a synthetic image (left) and corresponding tree of shapes (right). Alphabet letters denote the correspondence between blobs and tree nodes, and numbers denote gray levels. Inequality signs, i.e., $<$ and $>$, denote dark and bright nodes, respectively.

4.2.1 Tree of Shapes

A tree of shapes [35, 109] is an efficient image representation in a self-dual form. Given an image $u : \mathbb{R}^2 \rightarrow \mathbb{R}$, the upper and lower level sets of u are defined for $\lambda \in \mathbb{R}$ as follows:

$$(4.1) \quad \chi_\lambda(u) = \{x \in \mathbb{R}^2 \mid u(x) \geq \lambda\}$$

$$(4.2) \quad \chi^\lambda(u) = \{x \in \mathbb{R}^2 \mid u(x) < \lambda\}.$$

From these level sets, we can obtain tree structures $\mathcal{T}_\geq(u)$ and $\mathcal{T}_<(u)$ that comprise connected components of upper- and lower-level sets as follows:

$$(4.3) \quad \mathcal{T}_\geq(u) = \{\Gamma \mid \Gamma \in \mathcal{CC}(\chi_\lambda(u)), \forall \lambda\}$$

$$(4.4) \quad \mathcal{T}_<(u) = \{\Gamma \mid \Gamma \in \mathcal{CC}(\chi^\lambda(u)), \forall \lambda\},$$

where $\mathcal{CC}$ is an operator giving a set of connected components.

Furthermore, we define a set of upper shapes $\mathcal{S}_\geq(u)$ and lower shapes $\mathcal{S}_<(u)$. These sets are obtained by the cavity filling (saturation) of components of $\mathcal{T}_\geq(u)$ and $\mathcal{T}_<(u)$. A *tree of shapes* of u is defined as the set of all shapes defined as $\mathcal{G}(u) = \mathcal{S}_\geq(u) \cup \mathcal{S}_<(u)$.

As a consequence of the nesting property of level sets, the tree of shapes forms a hierarchical structure. Figure 4.2 shows an example of a tree of shapes. Let $T = \{V, E\}$ be a tree of shapes, where $V = \{v_j\}$ is a set of nodes and $E = \{(v_j, v_k)\}$ is a set of edges. Let s_j be a blob in u corresponding to v_j , and let a_j be the area (the number of pixels) of s_j . The area of an image u is denoted as A . We define parent and children nodes of v_j as

$$(4.5) \quad Pa(v_j) = \{v_k \mid (v_j, v_k) \in E, a_j < a_k\}$$

$$(4.6) \quad Ch(v_j) = \{v_k \mid (v_j, v_k) \in E, a_j > a_k\},$$

respectively, and we similarly define $Pa(s_j)$ and $Ch(s_j)$. Note that we use blob s_j and node v_j interchangeably in the following discussion.

4.2.2 SITA

Xia et al. [158] proposed SITA, a texture descriptor based on the tree of shapes. It comprises four features of blobs corresponding to nodes.

First, the $(p + q)$ th order central moment μ_{pq} of s_j is defined by

$$(4.7) \quad \mu_{pq}(s_j) = \int \int_{s_j} (x_j - \bar{x}_j)^p (y_j - \bar{y}_j)^q dx_j dy_j,$$

where $(\bar{x}_j, \bar{y}_j)$ is the center of mass of s_j . The normalized moments are defined as

$$(4.8) \quad \eta_{pq}(s_j) = \frac{\mu_{pq}(s_j)}{\mu_{00}(s_j)^{(p+q+2)/2}}.$$

Then, two eigenvalues, $\lambda_{1j}, \lambda_{2j}$, ($\lambda_{1j} \geq \lambda_{2j}$), of the normalized inertia matrix

$$(4.9) \quad C(s_j) = \begin{pmatrix} \eta_{20}(s_j) & \eta_{11}(s_j) \\ \eta_{11}(s_j) & \eta_{02}(s_j) \end{pmatrix}$$

are computed.

The first two features of SITA are elongation

$$(4.10) \quad \epsilon(s_j) = \frac{\lambda_{2j}}{\lambda_{1j}}$$

and compactness

$$(4.11) \quad \kappa(s_j) = \frac{1}{4\pi\sqrt{\lambda_{1j}\lambda_{2j}}}.$$

The third feature is the scale ratio $\alpha(s_j)$ defined by

$$(4.12) \quad \alpha(s_j) = \frac{\mu_{00}(s_j)}{\sum_{s_k \in \cup_M Pa^M(s_j)} \mu_{00}(s_k) / M},$$

where $Pa^M(s_j)$ is the M th ancestor blob defined by

$$(4.13) \quad Pa^1(s_j) = Pa(s_j)$$

$$(4.14) \quad Pa^2(s_j) = Pa^1(Pa^1(s_j))$$

$$\vdots$$

$$(4.15) \quad Pa^M(s_j) = Pa^1(Pa^{M-1}(s_j))$$

This is the ratio of blob sizes between s_j and the ancestor blobs. In accordance with [158], we set $M = 3$ in our experiments.

The fourth feature comprises normalized gray values, $\{\gamma(x)\}$, computed for each pixel $x \in s_j$ as follows:

$$(4.16) \quad \gamma(x) = \frac{u(x) - m_{j(x)}}{\sigma_{j(x)}},$$

where $m_{j(x)}$ and $\sigma_{j(x)}^2$ are the mean and variance of $u(x)$ over $s_{j(x)}$, respectively.

$$(4.17) \quad m_{j(x)} = \frac{1}{a_{j(x)}} \sum_{x \in s_{j(x)}} u(x)$$

$$(4.18) \quad \sigma_{j(x)}^2 = \frac{1}{a_{j(x)}} \sum_{x \in s_{j(x)}} (u(x) - m_{j(x)})^2.$$

Here, $s_{j(x)}$ is the smallest blob containing x , where $j(x) = \arg \min_j \{a_j | x \in s_j\}$.

Then, the computed four texture features are used for histogram feature computation with respect to dark and bright nodes. Here, let $v(s_j)$ be the gray level of blob s_j defined as follows:

$$(4.19) \quad v(s_j) = \frac{1}{\mu_{00}(s_j) - \mu_{00}(Ch(s_j))} \sum_{x \in s_j / Ch(s_j)} u(x).$$

A blob is defined as dark if $v(s_j) \leq v(Pa(s_j))$ and as bright otherwise. The root node is simply defined as dark. For all dark nodes, we compute three histograms of the first three texture features, i.e., $\epsilon(s_j)$, $\kappa(s_j)$, and $\alpha(s_j)$, and we do the same for all bright nodes. For all nodes, we compute a histogram of $\{\gamma(x)\}$. Consequently, we obtain seven histograms¹. These seven histograms are concatenated into a single one, which is called a SITA feature.

4.2.3 Recursive Representation of SITA

In the original study, the authors did not describe how to compute a SITA feature. Here, we propose a recursive procedure because of the relation to our proposed method. The SITA computation can be performed by aggregating histograms from leaf nodes to the root as follows. Let $\mathbf{g}_j$ be a concatenated histogram computed from node v_j only. The aggregated histogram $\mathbf{h}(v_j)$ from nodes below v_j in the tree is computed recursively as

$$(4.20) \quad \mathbf{h}_j = \mathbf{g}_j + w_{agg} \sum_{v_k \in Ch(v_j)} \mathbf{h}_k,$$

where w_{agg} is the weight for aggregating histograms of children nodes. Finally, the histogram $\mathbf{h}_{root}$ at the root node v_{root} is normalized to have a unit L_1 norm with respect to each of the seven histograms.

4.3 Proposed Method

Here, we define notions of trees of shapes for a set of images. Let $\{u_i\}_{i=1}^N$ be a set of images and A_i be the area of u_i . A tree of shapes of u_i is defined as $T_i = \{V_i, E_i\}$, and $n_i = |V_i|$ is the number of nodes in T_i . Each node $v_{ij} \in V_i$ has the corresponding blob s_{ij} with the area a_{ij} .

¹In this study, we set the number of bins as 25 and the histogram range as (0, 1) for $\epsilon(s_j)$, $\kappa(s_j)$, and $\alpha(s_j)$. For $\{\gamma(x)\}$, we set the number of bins as 50 and the histogram range as (-25, 10). Note that we set the histogram range for $\{\gamma(x)\}$ experimentally.

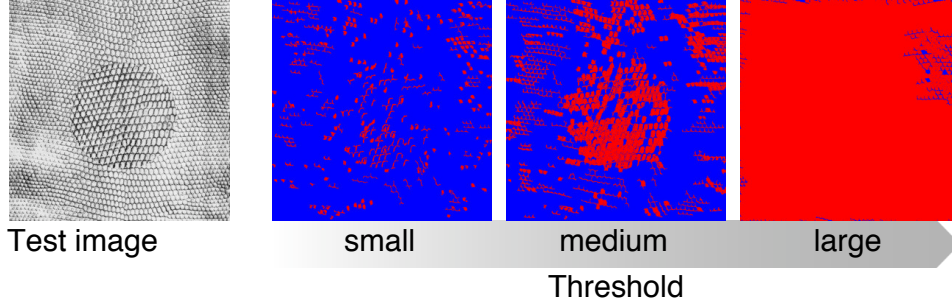


Figure 4.3: Examples of labeling results using subtrees with different node sizes. Subtrees used for labeling are determined using an estimated threshold.

We assume that each blob is given a ground truth label $y_{ij} \in L$ for training images, where L is a set of labels (in our case $L = \{-1, 1\}$.)

In contrast to the original SITA, we compute aggregated histograms at *every* node. Let $\mathbf{g}_{ij}$ be a histogram computed from node v_{ij} only. Then, the aggregated histogram $\mathbf{h}_{ij}$ nodes below node v_{ij} in the tree are computed recursively as

$$(4.21) \quad \mathbf{h}_{ij} = \mathbf{g}_{ij} + w_{agg} \sum_{v_{ik} \in Ch(v_{ij})} \mathbf{h}_{ik},$$

and then normalized to have a unit L_1 norm.

We mainly compute SITA features at root nodes of *all* subtrees in the tree, whereas the original SITA features are computed at the root node only. One of the simple ideas for image labeling is to classify these node-wise SITA features to obtain labels of blobs. As mentioned previously, small subtrees are not useful for classification. Figure 4.3 shows examples of unstable labeling results. Here, we set three different thresholds to areas of subtree root nodes and classify SITA features of the subtrees that are larger than the thresholds for labeling. As shown in the figure, the results would not be satisfactory with excessively small (or large) subtrees with an excessively small (or large) area threshold. Based on this observation, we assume that there exists an optimal threshold for the area (or size) of subtrees. Furthermore, we have no reason to expect that a single area threshold is desirable for different training images whose texture contents might be different.

Therefore, we formulate the task as a joint optimization problem, estimating thresholds for each training image and training a classifier. Let θ_i be a threshold for u_i , and let $\mathbf{w}$ and b be parameters of a classifier (here, using SVMs, these are the weight vector and the bias, respectively). We define the objective function for u_i as follows:

$$(4.22) \quad E_i(\theta_i, \mathbf{w}, b) = \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{n_i} \sum_j^{n_i} W_{ij} \ell(y_{ij}(\mathbf{w}^T \mathbf{h}_{ij} + b)) + \frac{\lambda}{2} \theta_i^2.$$

The first term is the SVM regularizer, and in the second term, $\ell(\cdot)$ is the hinge loss function of the SVM. The third term is the regularizer for θ_i , and λ is the scale parameter.

In the objective function, we introduce the sample weight W_{ij} for $\mathbf{h}_{ij}$. In the proposed method, we use θ_i to threshold smaller subtrees. In other words, we use the histograms of subtrees larger than θ_i and ignore the others. This is the basic concept, and it can be implemented by setting zero or one as values of W_{ij} . However, this is difficult to solve as an optimization with gradient-based solvers. Therefore, we adopt a sigmoid function for representing the thresholding and define W_{ij} as follows:

$$(4.23) \quad W_{ij} = W(a_{ij}, \theta_i) = \frac{1}{1 + e^{-\beta(a_{ij} - \theta_i)}},$$

where β is the gain parameter of the sigmoid.

In the training phase, we have a set of N training images $\{u_i\}_{i=1}^N$, and the objective function to be minimized for training is

$$(4.24) \quad \begin{aligned} E(\boldsymbol{\theta}, \mathbf{w}, b) &= \frac{1}{N} \sum_i^N E_i(\theta_i, \mathbf{w}, b) \\ &= \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{N} \sum_i^N \frac{1}{n_i} \sum_j^{n_i} W(a_{ij}, \theta_i) \ell(y_{ij}(\mathbf{w}^T \mathbf{h}_{ij} + b)) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2, \end{aligned}$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_N)^T$.

4.3.1 Optimization

Given a training set of images, we estimate parameters $\hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}, \hat{b}$ by

$$(4.25) \quad \hat{\boldsymbol{\theta}}, \hat{\mathbf{w}}, \hat{b} = \underset{\boldsymbol{\theta}, \mathbf{w}, b}{\operatorname{argmin}} E(\boldsymbol{\theta}, \mathbf{w}, b).$$

Since this is non-linear and non-convex, we use a block-coordinate decent approach, that is, given initial value $\boldsymbol{\theta}_0$, we iteratively estimate, first, the classifier parameters $\mathbf{w}$ and b and then the thresholds $\boldsymbol{\theta}$.

4.3.1.1 Classifier Training

To estimate $\mathbf{w}$ and b , given $\boldsymbol{\theta}_{k-1}$, we solve

$$(4.26) \quad \mathbf{w}_k, b_k = \underset{\mathbf{w}, b}{\operatorname{argmin}} E(\boldsymbol{\theta}_{k-1}, \mathbf{w}, b).$$

This is an SVM formulation with sample weights, which is convex. We solve this problem using the primal solver of LIBLINEAR [29] because dual solvers are difficult to apply to a large number of training samples. (In our case, the SVM is trained on approximately hundred thousand node features.)

4.3.1.2 Threshold Estimation

To estimate θ , given $\mathbf{w}_k$ and b_k , we solve

$$(4.27) \quad \theta_k = \underset{\theta}{\operatorname{argmin}} E(\theta, \mathbf{w}_k, b_k).$$

This is non-convex because θ depends on histograms $\mathbf{h}_{ij}$. We solve this using Newton's method because we confirmed experimentally that the cost function is smooth and has a single minimum in many cases (details are described in Section 4.4.1), and the Hessian is diagonal, as shown below.

The gradient is given by

$$(4.28) \quad \nabla E = \left(\frac{\partial E}{\partial \theta_1}, \dots, \frac{\partial E}{\partial \theta_N} \right),$$

where

$$(4.29) \quad \frac{\partial E}{\partial \theta_i} = \frac{1}{N n_i} \sum_j^{n_i} -\beta W_{ij} (1 - W_{ij}) \ell(y_{ij}(\mathbf{w}^T \mathbf{h}_{ij} + b)) + \lambda \theta_i.$$

The Hessian is

$$(4.30) \quad \nabla^2 E = \begin{pmatrix} \frac{\partial^2 E}{\partial \theta_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{\partial^2 E}{\partial \theta_N^2} \end{pmatrix},$$

where

$$(4.31) \quad \frac{\partial^2 E}{\partial \theta_i^2} = \frac{1}{N n_i} \sum_j^{n_i} \beta^2 W_{ij} (1 - W_{ij}) (1 - 2W_{ij}) \ell(y_{ij}(\mathbf{w}^T \mathbf{h}_{ij} + b) + \lambda)$$

The Hessian is diagonal because there are no cross terms in the second order derivatives. Therefore, we can parallelize the implementation to reduce the computation time.

4.3.1.3 Stopping Criterion

We stop the alternation when θ_k converges with the termination criterion of

$$(4.32) \quad \|\theta_k - \theta_{k-1}\| = \epsilon.$$

4.3.2 Labeling Procedure

Typically, in the labeling phase of a test image u , we first construct the tree of shapes of u , compute $\mathbf{h}_j$ for every node v_j , and then conceptually classify those $\mathbf{h}_j$ whose area a_j is larger than a threshold. However, here we choose to do this differently because we have estimated a set of thresholds θ_i for training images u_i . Instead of the above approach, we propose minimizing the objective function Eq. (4.22) again for the test image as we did in the training phase.

Algorithm 2 Labeling procedure

```

1: Input: threshold  $\hat{\theta}$ 
2: Input: SVM parameters  $\mathbf{w}$  and  $b$ 
3: Input: tree of shapes  $T$  of test image  $u$ 
4: Output: labeling result  $u_l$ 
5: initialize  $u_l = \mathbf{0}$ 
6: for node  $j = 1 \dots n$  do
7:   if  $a_j \geq \hat{\theta}$  then
8:      $y_j \leftarrow \text{sign}(\mathbf{w}^T \mathbf{h}_j + b) \setminus \setminus$  classify a histogram.
9:   else  $\{a_j < \hat{\theta}\}$ 
10:     $y_j \leftarrow y_{Pa(v_j)} \setminus \setminus$  assign label of parent node.
11:   end if
12: end for
13:  $u_l(x) = y_{j(x)}, j(x) = \text{argmin}_j \{a_j | x \in s_j\} \setminus \setminus$  map label to the corresponding pixel.
14: return  $u_l$ 

```

First, we fix the classifier parameters (hence, the loss in the objective function) and minimize the following objective function to estimate $\hat{\theta}$ as follows:

$$(4.33) \quad E(\theta, \mathbf{y}) = \frac{1}{2} \|\mathbf{w}\|^2 + \frac{1}{n} \sum_j^n W(a_j, \theta) \ell(y_j(\mathbf{w}^T \mathbf{h}_j + b)) + \frac{\lambda}{2} \theta^2,$$

where $\mathbf{y} = \{y_j\}_{j=1}^n$ is a set of labels. This is the same as with the threshold estimation in the training phase but for only a single test image (i.e., $N = 1$).

Algorithm 2 provides details of the labeling procedure. To obtain a segmentation result, we perform the classification procedure starting from the root node and proceeding down to the leaf nodes. At each node n_j , if $a_j \geq \hat{\theta}$, we classify $\mathbf{h}_j$ and then assign the resulting y_j to all pixels in s_j , even including those that have been assigned labels by parent nodes (i.e., overwriting labels). This downward traversal of the tree stops once it reaches smaller nodes.

To refine the segmentation result, we apply simple morphological filtering [135] as a post process.

4.4 Experimental Results

We tested the proposed method and compared it with existing methods using synthetic and real image datasets.

For the experiments, we created a synthetic dataset from the UIUC database [87], which comprises 25 various texture classes, each of which contains 40 images of size 640×480 pixels. This dataset comprises seven sub-datasets. Each sub-dataset includes five images containing one to three regions made of two classes. Figure 4.4 shows an example of a created sub-dataset containing five images and corresponding ground truth labels. Trees of shapes of these images have sufficient nodes, 18,855 nodes per image, on average. For each sub-dataset, we randomly select three images for training and two for testing. In the experiments, we set sigmoid gain $\beta = 0.1$, initial values of thresholds to 1,000, and weight for aggregating histogram $w_{agg} = 1.0$. For quantitative evaluation, we use the Dice coefficient [25].

We used three methods for comparison: Felsenszwalb’s unsupervised image segmentation [32], a patch-based MRF segmentation with SVM introduced in chapter 3, and a fully connected CRF [80] with TextonBoost [137, 138].

4.4.1 Energy Convergence

First, we show the convergence property of our proposed method. As mentioned previously, our method minimizes the cost function using Newton’s method for θ and SVM training for w, b . Here, we focus on the convergence of Newton’s method for the nonlinear optimization of θ because SVM training is convex and guaranteed to converge. Figure 4.5 shows the cost function values over different initial values and scale parameters λ for a training image. Note that we did not add the regularizer of SVM, i.e., $\|w\|^2$, to the cost function because it takes excessively large values that interfere with quantitatively observing the plot. In the following figures for cost function values, we also do not add the regularizer of SVM. As can be observed, the cost function is not convex,

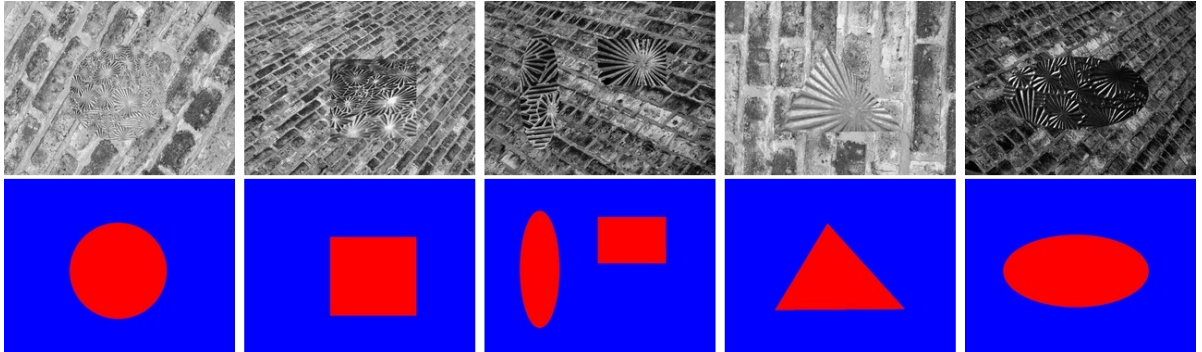


Figure 4.4: Example of a texture image sub-dataset. The upper and bottom rows show created texture images and corresponding ground truths, respectively. Blue and red indicate the class of each pixel.

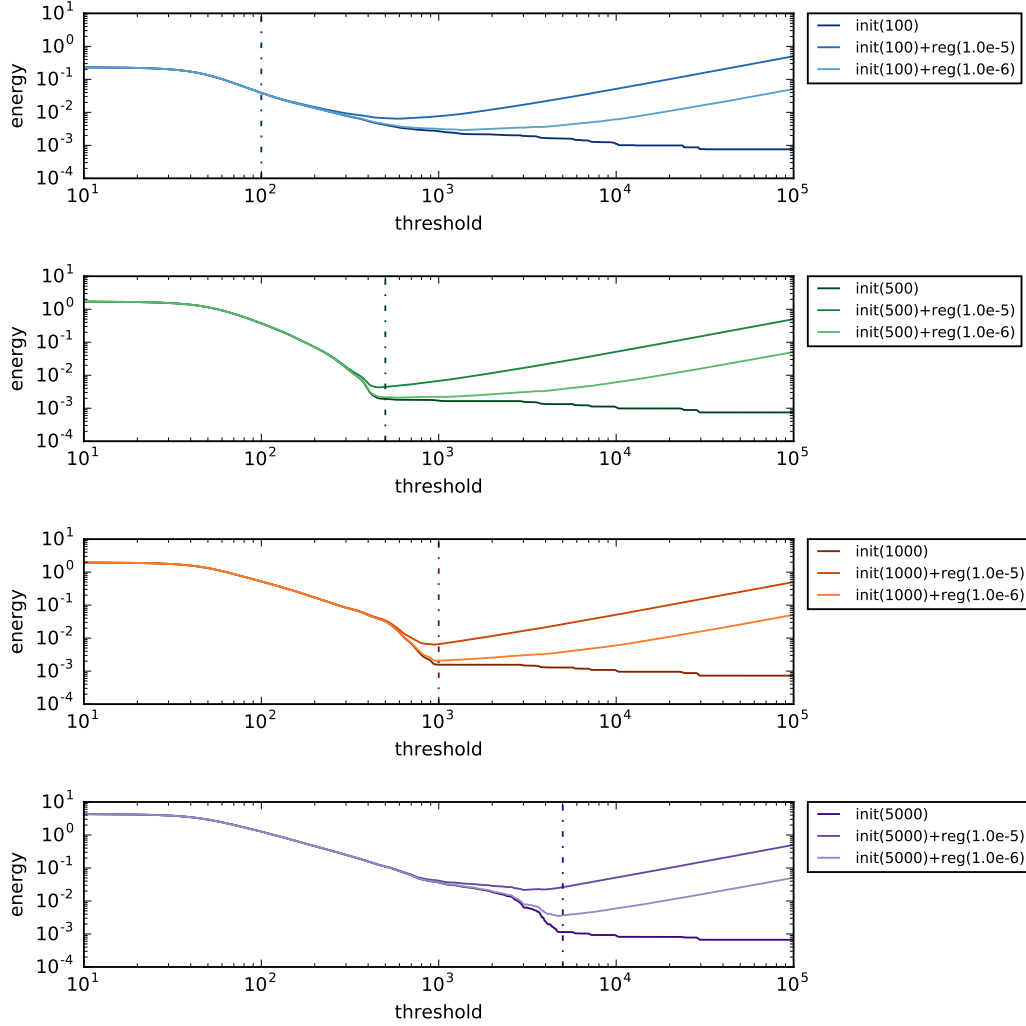


Figure 4.5: Cost function over different initial values and thresholds for a training image. The horizontal and vertical axes show the threshold and corresponding cost function value, respectively. Different colors are used for different initial values of θ_i . From top to bottom, the initial thresholds are 100, 500, 1,000, and 5,000.

but adding the regularizer for θ renders the energy values rather convex. Figure 4.6 shows the cost function values against the threshold over different iterations, as well as for different initial values θ_0 . Note that this figure shows the cost function values of a test image in the labeling phase because in the training phase we estimate θ_i separately for each training image, and it is difficult to visualize all of them in a single plot. We observe that the minimum of the cost function decreases, and the threshold θ_i converges with different initial values. However, it should be avoided to use a small initial value, such as $\theta_0 = 10$, because the cost function between $\theta = [10^0, 10^1]$ looks almost flat and non-convex and the iteration might not converge. It would be better to use a large initial value, typically larger than 1000. The bottom two plots show the

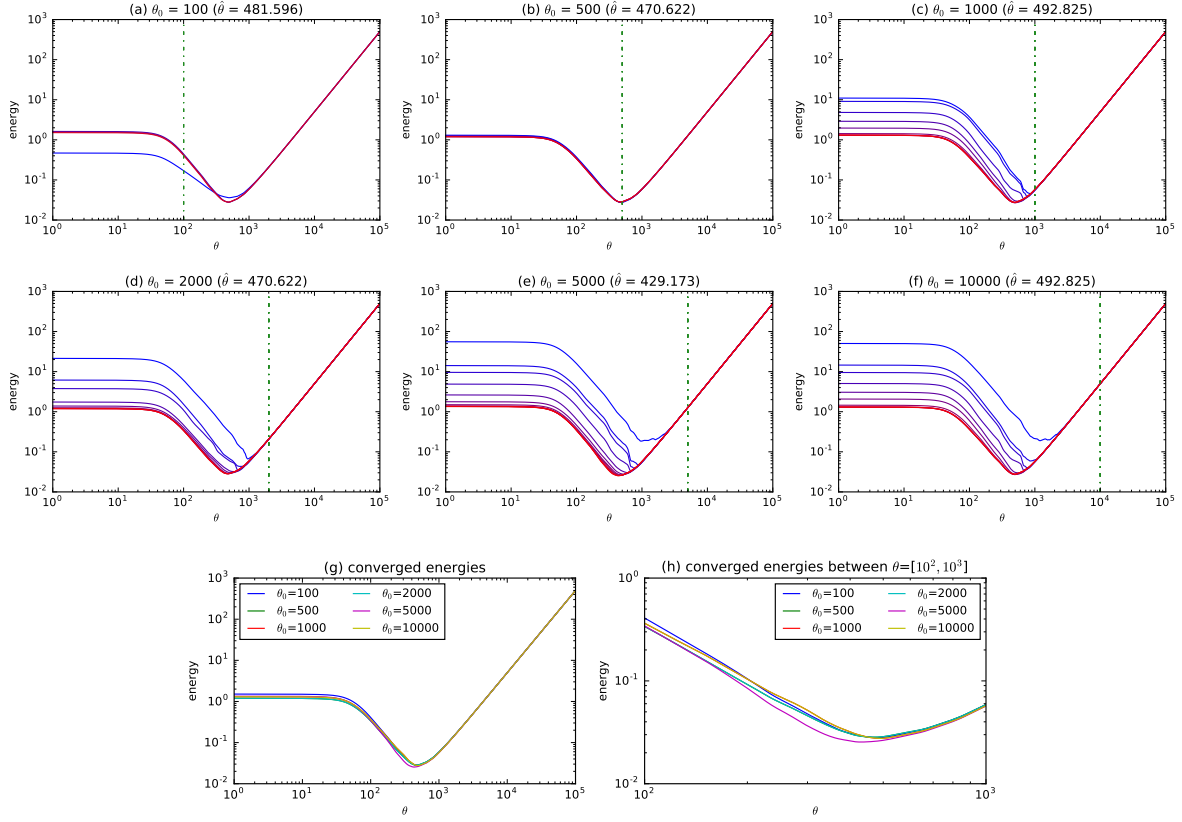


Figure 4.6: Cost function values against the threshold over different iterations, starting with different initial values $\theta_0 = 100, 500, 1000, 2000, 5000$, and 10000 . The horizontal and vertical axes show the threshold and corresponding cost function value, respectively. (a – f) Convergence with different initial values indicated as vertical green lines. Different colors are used for different iterations; the first iteration is in blue, and the final one is in red. (g) Final iterations taken from (a) to (f), and (h) its magnified version in the range of 10^2 and 10^3 .

cost function values of the final iterations of different initial values. The estimated thresholds $\hat{\theta}_i$ with different initial values are close to each other, however, the number of iterations and computation cost increase when a large initial value is used. Therefore, using too large values are not recommended. Figure 4.7 shows the cost function values and estimated thresholds of Figure 4.6(c) to show the convergence of the entire optimization procedure over iterations. We observe that the energy and threshold converge appropriately.

Next, we show the cost function values with different scale parameter values λ in Figure 4.8. The top-left plot $\lambda = 10^{-1}$ indicates that the value is too large so that the cost function is over-regularized and only the trivial estimates was obtained. As λ getting smaller, minimum becomes prominent and the estimated threshold shifts toward larger values. Labeling results with different λ values will be shown in Figure 4.11 in the following section.

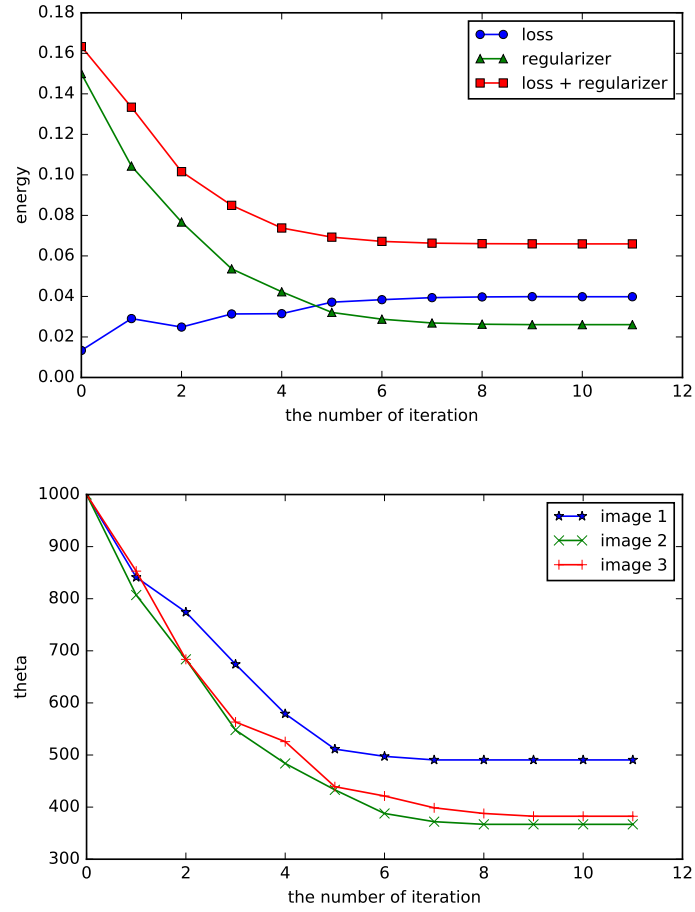


Figure 4.7: Energies (right) and thresholds (left) at each iteration. The horizontal and vertical axes show the number of iterations and the energy (right) and estimated threshold (left), respectively.

4.4.2 Labeling Results on a Synthetic Dataset

Figures 4.9 show the results for a synthetic sub-dataset. In Felzenszwalb's segmentation, there are too many boundaries that do not fit the ground truth label. The results of SVM-MRF and CRF are not qualitatively and quantitatively better than the results of the proposed method. The performances of the MRF- and CRF-based approaches are highly dependent on the unary term. In other words, failures by SVM and TextonBoost have too much impact on performance. For this kind of small dataset, MRF- and CRF-based methods are not the best choice for achieving good performance. In contrast, the proposed method gives reasonable labeling results and better performance in terms of the Dice coefficient. Note that since 42,169 nodes (or samples) are used for training, the primal solver for SVM training is necessary. Figure 4.10 shows results for another sub-dataset shown in Figure 4.4 that contains large geometrical, scale, and contrast changes. In this result, three low-contrast images are used for training, and the remaining ones are used for testing. MRF and CRF fail, whereas the proposed method labels test images reasonably. Figure

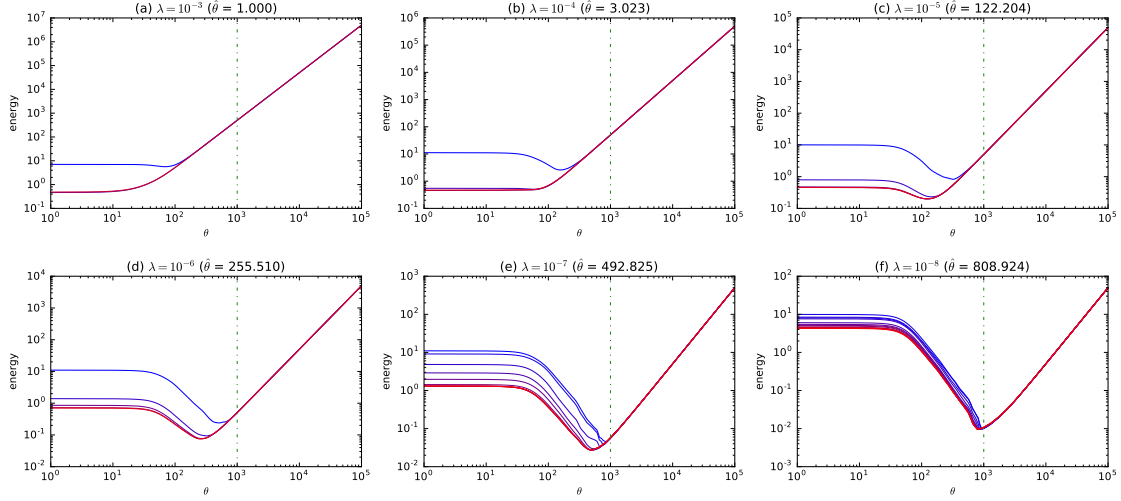


Figure 4.8: Cost function values against the threshold over different iterations with different scale parameter values $\lambda = 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}$, and 10^{-8} . The horizontal and vertical axes show the threshold and corresponding cost function value, respectively. (a – f) Convergence with different scale parameter values. The initial value $\theta_0 = 1000$ is indicated as vertical green lines. Different colors are used for different iterations; the first iteration is in blue, and the final one is in red.

4.11 shows labeling results of the proposed method with different scale parameter values λ . A large value of $\lambda = 10^{-3}$ provides a small threshold value, and the boundary between foreground and background disappear. Smaller values $\lambda = 10^{-7}$ and 10^{-8} , provide larger threshold values, and foreground objects becomes smaller. A better labeling result can be obtained when an appropriate value of λ , in this case 10^{-5} . The value should be tuned as the accuracy of labeling results is sensitive to it. Results for other sub-datasets are shown in Figure 4.12 to 4.14. In these results, MRF and CRF provide some successful results but the proposed method is stable and better in most of the cases.

For quantitative evaluation, we compute Dice coefficients over different numbers of training images. In this experiment, we used a sub-dataset containing ten images by adding five images to the sub-dataset of Figure 4.4. Figure 4.15 shows the box plots of Dice coefficients for different numbers of training images. The overall Dice coefficients of MRF and CRF are lower than those of the proposed method. Even when nine images are used for training, the median of the Dice coefficients of MRF and CRF are approximately 0.6. Some of the Dice coefficients of CRF are extremely low. In contrast, the proposed method works effectively and provides adequately high Dice coefficients, even when only one training image was used.

Here, we show some failure cases of the proposed method, such as the examples shown in Figure 4.16. In these results, CRF outperforms the other methods. Two textures in the sub-dataset are of pebbles that are similar to each other, and the SITA feature used in the proposed method is invariant to this difference between geometrical, scale, and contrast changes. Therefore,

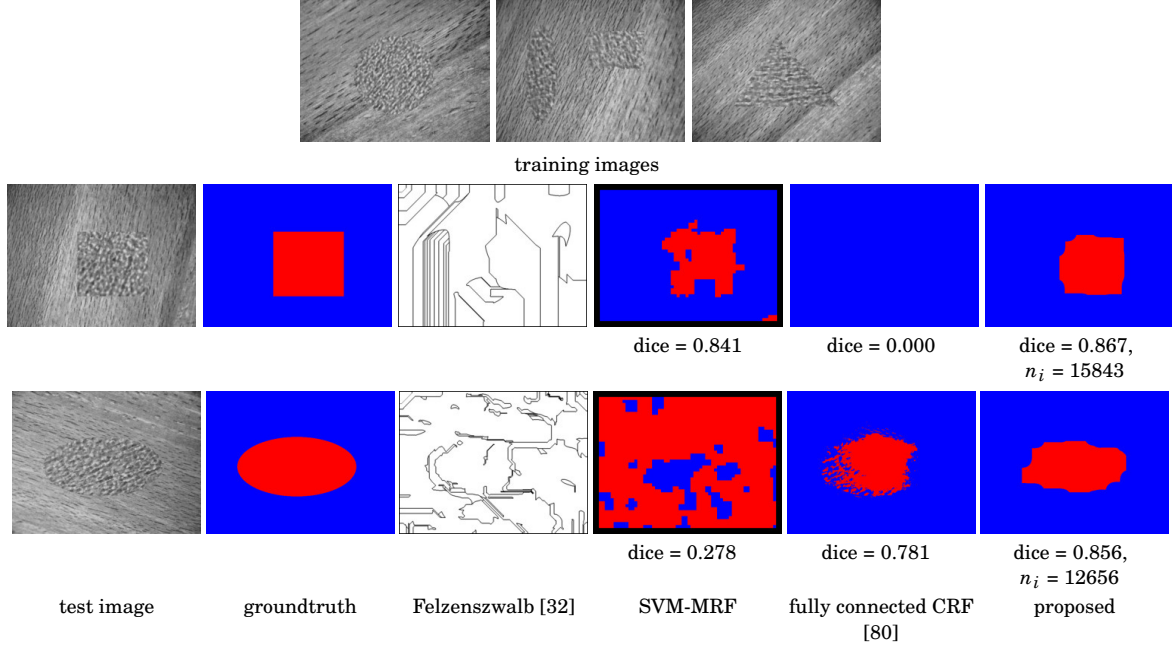


Figure 4.9: Labeling results. The Dice coefficient and number of nodes n_i are shown below the images. Red and blue represent each texture class, and the black of the SVM-MRF results represents a region that is unlabeled as a result of the boundary effect. λ is set to 10^{-6} , and the number of training samples is 42,169.

the subtrees cannot be classified correctly, and we obtained poor labeling results.

Figure 4.17 shows further results wherein all of the methods do not work well. The possible reason for this failure is that the number of nodes n_i is relatively small compared to that used for the successful results, such as in those in Figure 4.9. In this result, we observe that the number of training samples or nodes in the tree of shapes must be greater than at least 10,000 per image in order to obtain satisfactory labeling results.

Regarding the computational time, our Python implementation of the proposed method takes approximately 200 s for training and approximately 30 s for labeling an image in the above experimental condition. Although we handle more than 10,000 training samples (or nodes), our method can be trained within a practical time.

4.4.3 Labeling Results on the MSRC-21 Dataset

Hereafter, we show labeling results on the MSRC-21 dataset [138]. This dataset consists of 591 images with ground truth labels of 21 object classes. It has 20 subsets according to main objects of the image shown in the center, such as cow, sheep, tree, car, and building. To evaluate the proposed method with a few training images, we selected two subsets having rich texture contents; subset 2 (tree, grass, and sky) and 9 (sheep and grass). In each subset, the label of the main object of the subset (trees of subset 2, and sheep of subset 9) is used as foreground, and the

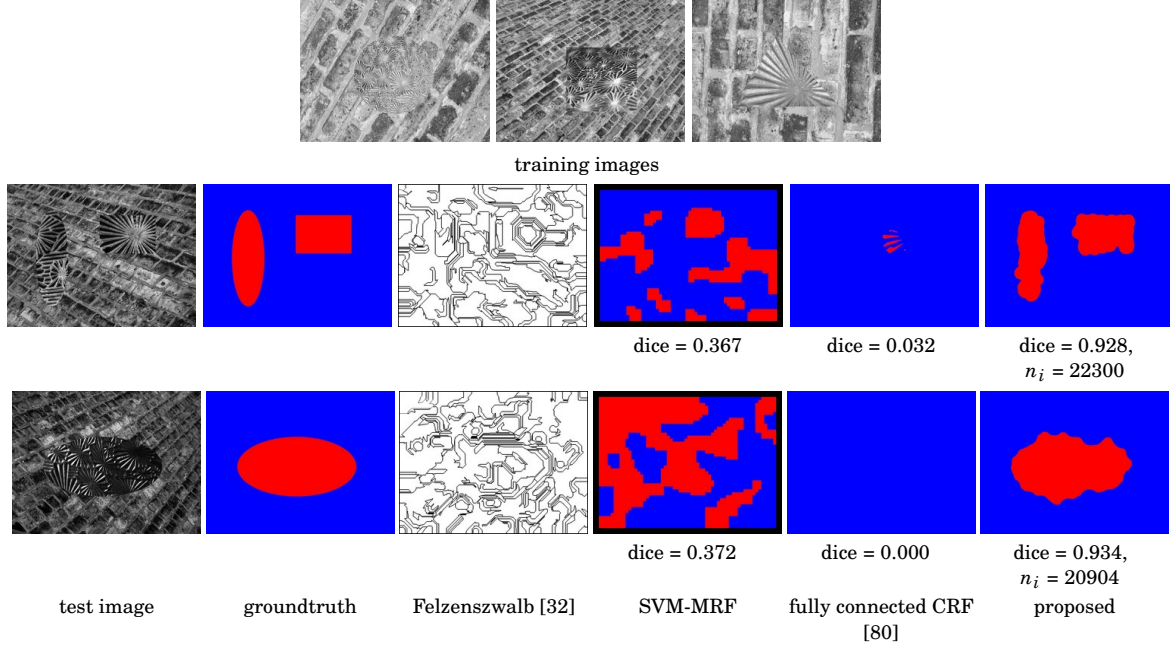


Figure 4.10: Labeling results. The Dice coefficient and number of nodes n_i are shown below the images. Red and blue represent each texture class, and the black of the SVM-MRF results represents a region that is unlabeled as a result of the boundary effect. λ is set to 10^{-5} , and the number of training samples is 59,701.

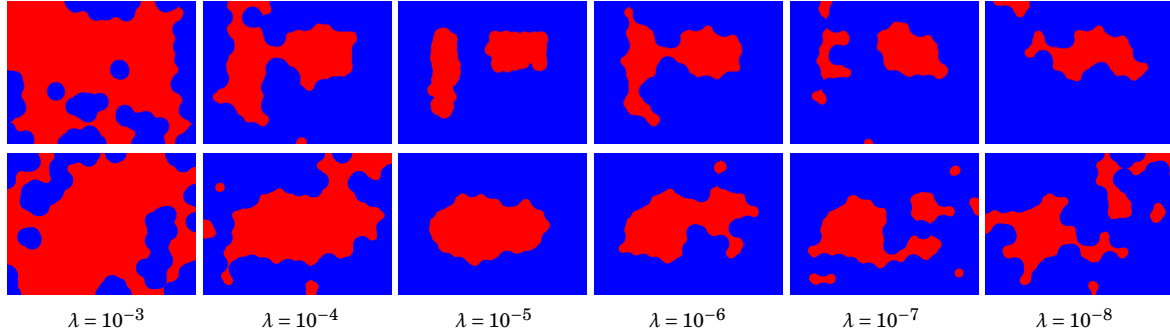


Figure 4.11: Labeling results with different scale parameter values λ , with the same images with Figure 4.10.

others are as background.

In natural images, color is an important cue for segmentation. To demonstrate the proposed method for color natural images in this experiment, we use the following tree of shapes and histogram features. First, we adopted the tree of shapes for color images [17]. The color version of the tree of shapes is constructed by merging a set of trees of shapes of each color component (in this paper, we used RGB) based on shapes (connected components) and their inclusion relationships. For more details, please refer to [17]. Moreover, we introduce two new histogram features in addition to the SITA. One is a histogram of HSV color values. We construct histograms of each

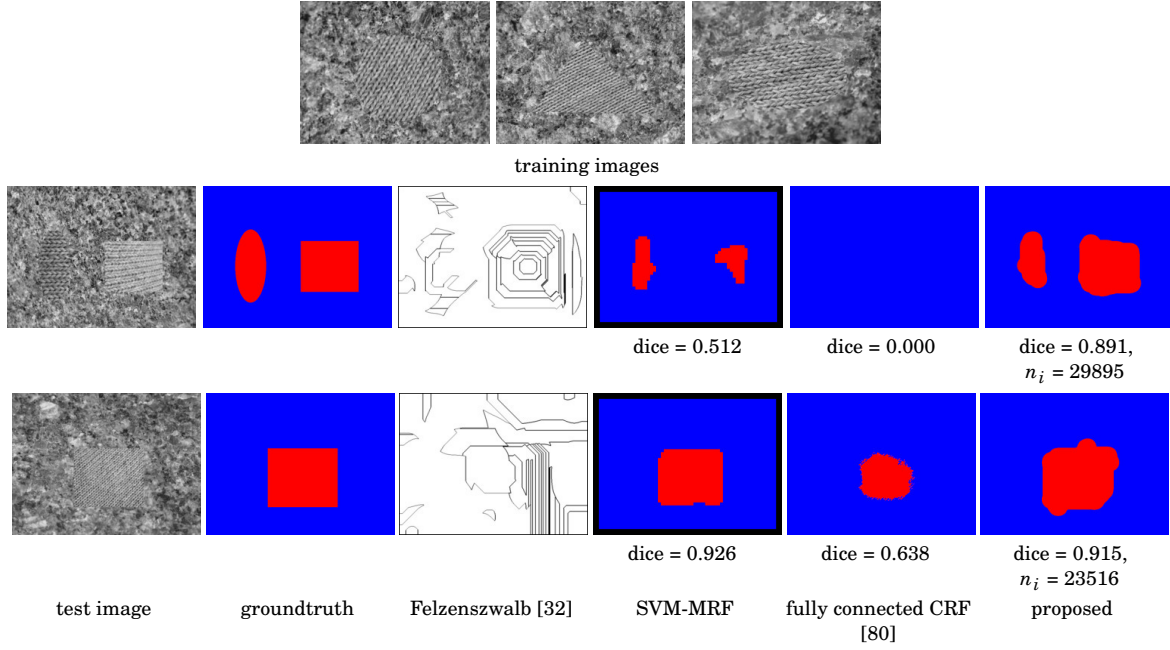


Figure 4.12: Labeling results. The Dice coefficient and the number of nodes n_i are shown below the images. Red and blue represent each texture class, and the black of the SVM-MRF results represents a region that is unlabeled as a result of the boundary effect. λ is set to 10^{-5} , and the number of training samples is 82,218.

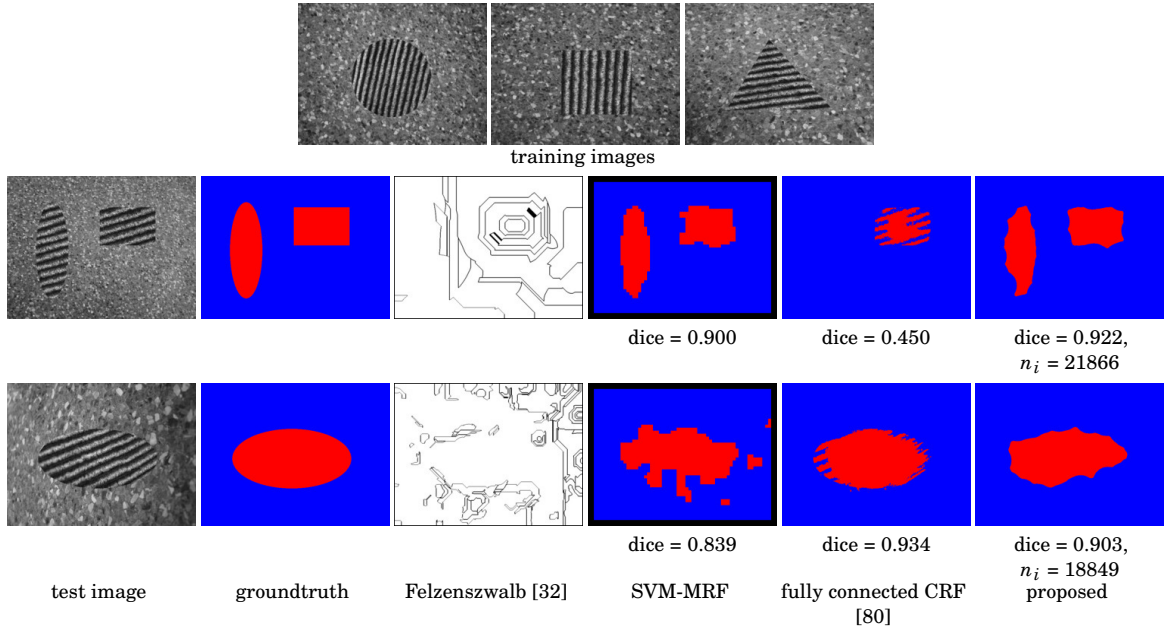


Figure 4.13: Labeling results. The Dice coefficient and the number of nodes n_i are shown below the images. Red and blue represent each texture class, and the black of the SVM-MRF results represents a region that is unlabeled as a result of the boundary effect. λ is set to 10^{-5} , and the number of training samples is 76,066.

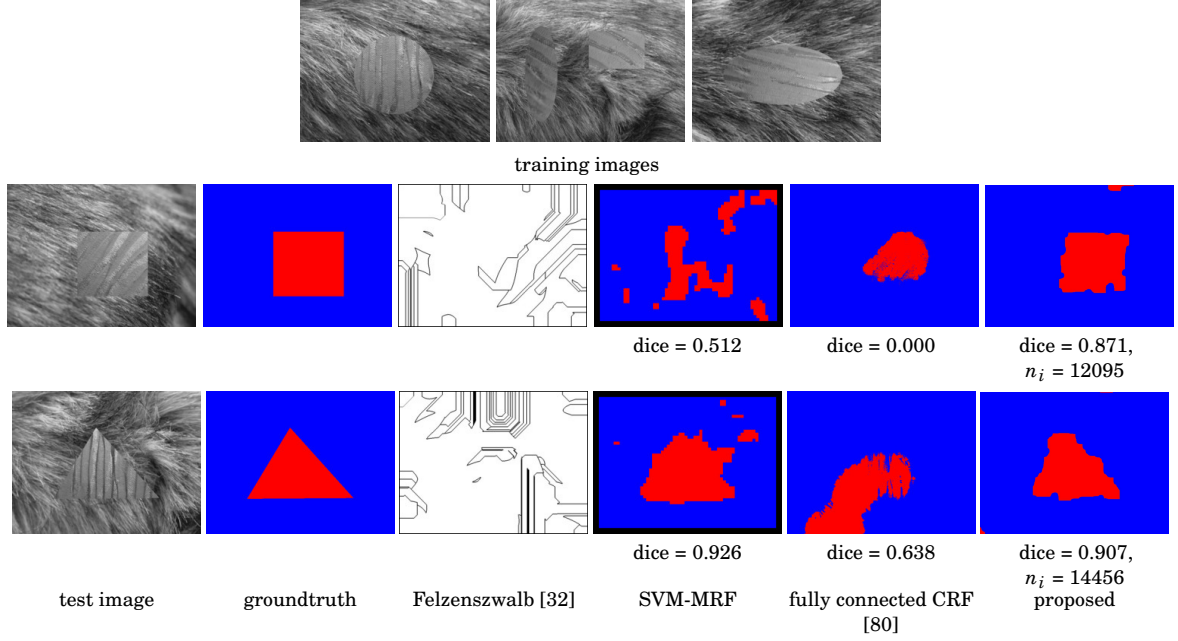


Figure 4.14: Labeling results. The Dice coefficient and the number of nodes n_i are shown below the images. Red and blue represent each texture class, and the black of the SVM-MRF results represents a region that is unlabeled as a result of the boundary effect. λ is set to 10^{-5} , and the number of training samples is 46,620.

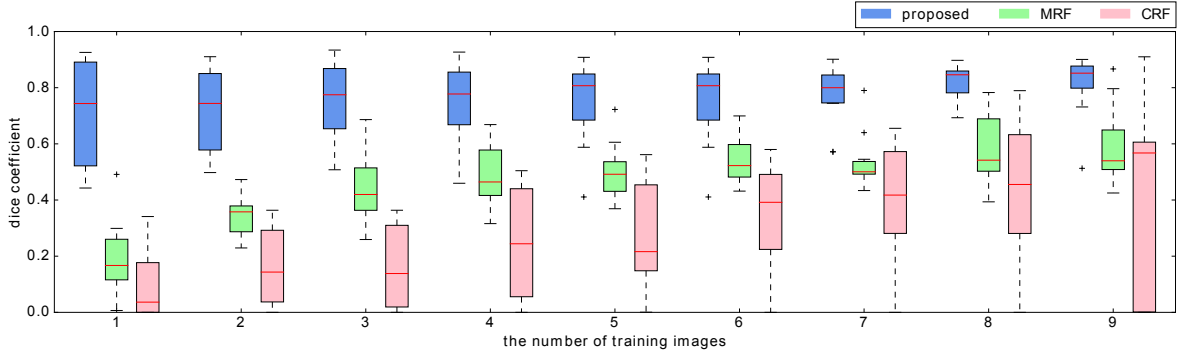


Figure 4.15: Box plots for Dice coefficients over different numbers of training images. The horizontal and vertical axes show the number of training samples and the Dice coefficients, respectively.

HSV component and concatenate to the original SITA histogram feature. The number of bins of each color component histogram is set to 50, then the total number of bins of the HSV histogram is 150. The other is a histogram of textons. We used 17 kernels used in the TextonBoost [138], and also 32 Gabor kernels². The 49-dimensional responses of training images are clustered by

²Used Gabor kernels consist of real and imaginary part of scales 3 and 5, frequencies 0.1 and 0.2, and rotations 0, $\pi/4$, $\pi/2$ and $3\pi/4$, which is decided experimentally. These Gabor kernels are convolved with the L component of the Lab color space.

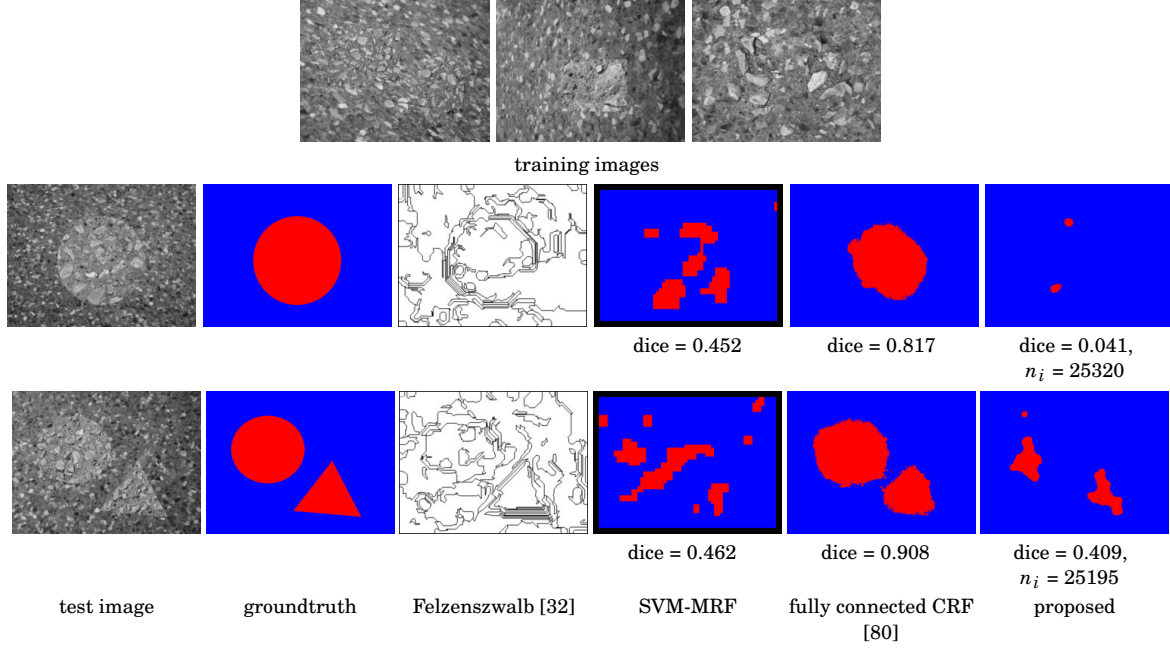


Figure 4.16: Some failure labeling results. The Dice coefficient and number of nodes n_i are shown below the images. Red and blue represent each texture class, and the black of the SVM-MRF results represents a region that is unlabeled as a result of the boundary effect. λ is set to 10^{-6} , and the number of training samples is 63,403.

the K-means algorithm. Then, each response is assigned to the nearest cluster center, or texton, and a histogram of these textons is created. The number of bins of the texton histogram is set to 200. In the experiments, we set sigmoid gain $\beta = 0.1$, initial values of thresholds to 1,000, and weight for aggregating histogram $w_{agg} = 0.8$. As a comparison, we used the fully connected CRF [80].

We show effect of the weight w_{agg} for aggregating children histograms. Figure 4.18 shows box plots for Dice coefficients over different values of w_{agg} . When $w_{agg} = 1.0$, dice coefficients are relatively lower than that of smaller w_{agg} values. With large values of w_{agg} , histograms are affected by features of small children nodes, while smaller values of w_{agg} result in less discriminative histogram features. In this experiment, we empirically set $w_{agg} = 0.8$.

Figures 4.19 and 4.20 show segmentation results on subset 2 and 9. These results are obtained with five training images ($N = 5$). CRF fails to classify pixels correctly due to the small number of training images. Meanwhile, better results are obtained by the proposed method.

Figure 4.21 shows box plots for Dice coefficients over different number of training images from each of two subsets. N training images are randomly selected from a subset, and then 10 test images are randomly selected from the rest of the subset. For subset 2 (top row), the proposed method performs better than CRF when fewer than 7 images are used. With 9 and 10 training images, CRF works better as expected because typically CRF needs many training images. For subset 9 (bottom row), the proposed method consistently outperform CRF even with 10 training

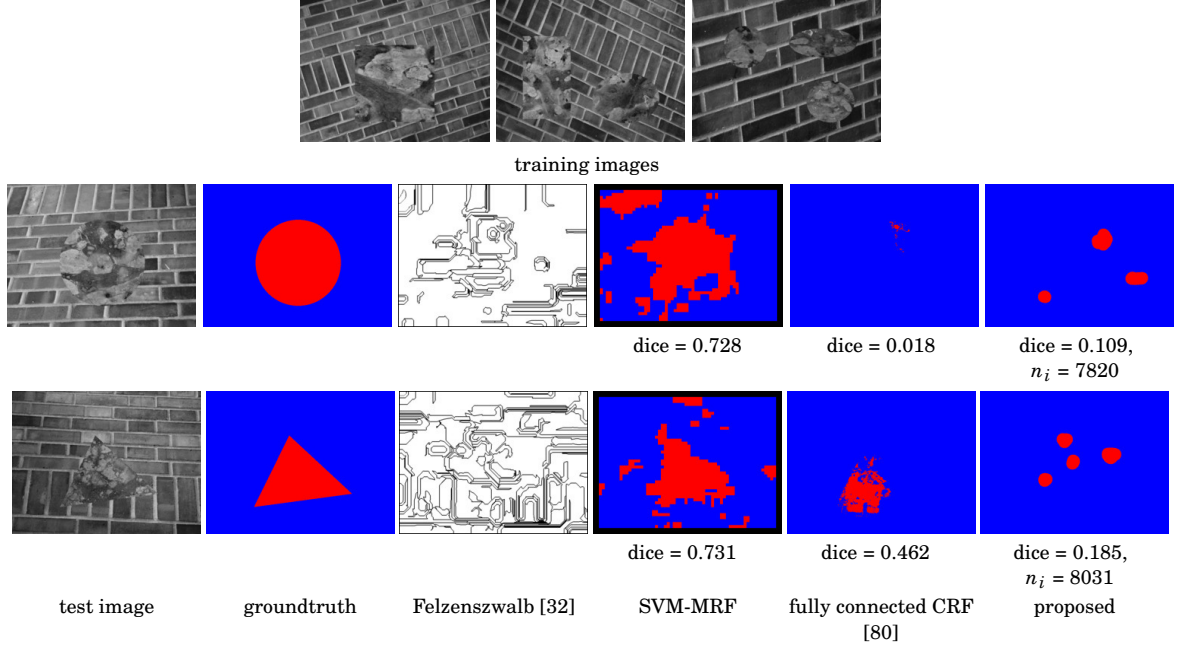


Figure 4.17: Some failure labeling results. The Dice coefficients and number of nodes n_i are shown below the images. Red and blue represent each texture class, and the black of the SVM-MRF results represents a region that is unlabeled as a result of the boundary effect. λ is set to 10^{-5} , and the number of training samples is 27,045.

images used.

4.4.4 Labeling Results on the NBI Endoscopic Images

We have prepared a dataset of 63 NBI endoscopic images. Example images in the dataset are shown in Figure 4.22. Sizes of images are $1,000 \times 870$ pixels. There are two label categories (foreground and background) based on the NBI magnification findings (see Figure 1.6). Foreground regions correspond to polyps of types B and C, and background regions are others (type A polyps, normal intestinal walls, and uninformative dark regions). Among 63 images, 20 images are negative samples which don't contain any foreground regions; the left-most image in Figure 4.22 captures only a hyperplastic polyp (i.e. benign tumor and non-cancer, hence Type A) labeled as background. A tree of shapes created from an NBI endoscopic image contains a large number of nodes. The average number of nodes from images in the dataset is 24,070. We randomly divided the dataset into half for training and test. We set parameters λ as 1.0 and initial value of threshold θ_0 as 1000.

We used two methods for comparison. One is to simply classify histograms of nodes in a tree of shapes and assign labels to pixels, which is corresponding to $W_{ij} = 1$ in Eq. (4.22). This is a simple application of SITA for every nodes and is an obvious extension, while our proposed method is not. In the following experiments, we refer this method as conventional method. The other

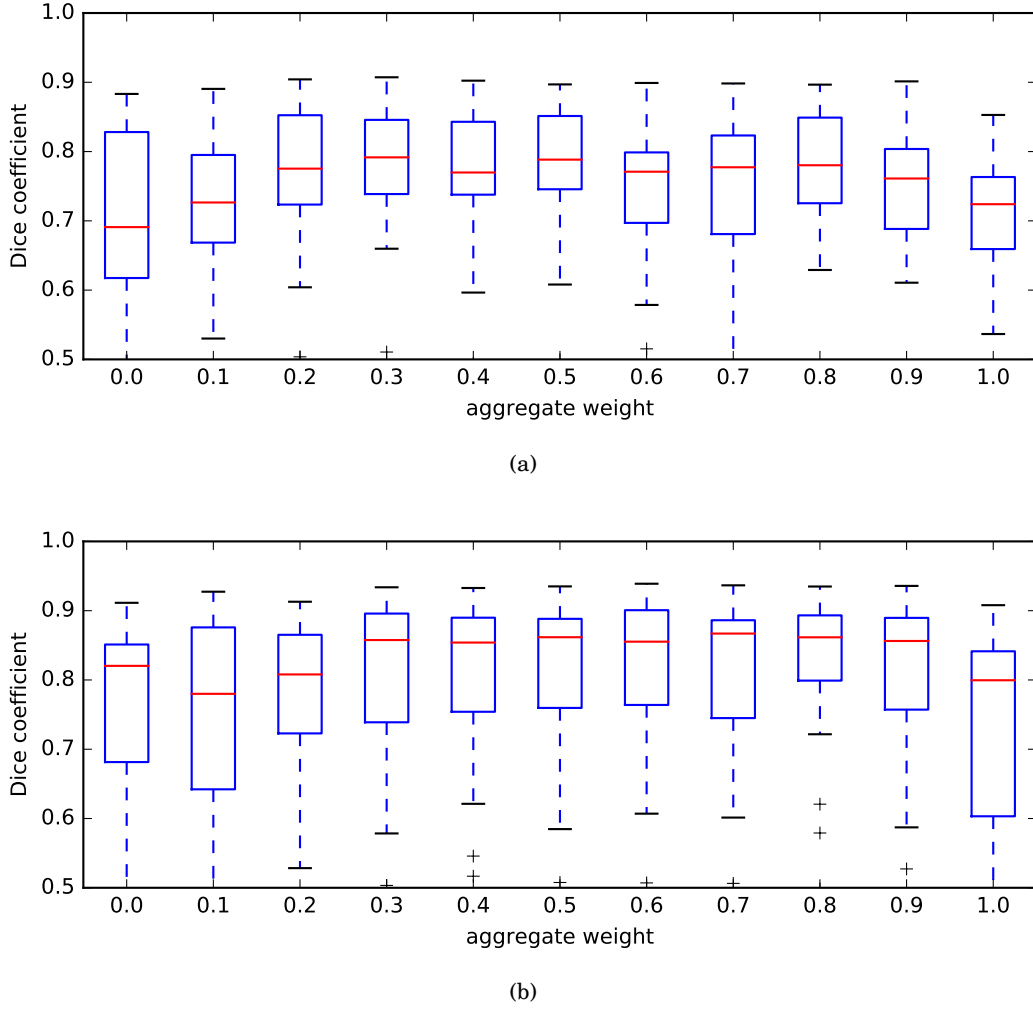


Figure 4.18: Box plots for Dice coefficients over different w_{agg} on (a) subset 2 and (b) subset 9 of the MSRC-21 dataset. The horizontal and vertical axes show w_{agg} and the Dice coefficients, respectively. The number of training images N is fixed to 5.

is a patch based segmentation method using MRF and posterior probabilities obtained from a trained SVM classifier introduced in chapter 3. For training SVM, we used 1,608 NBI endoscopic image patches (type A: 484, types B and C3: 1,124) trimmed and labeled by endoscopists. In this method, densely sampled SIFT features are extracted from these patches and converted as BoVW histograms. BoVW histograms are then used for training an SVM classifier. Small square patches corresponding to each site of the MRF grid are classified to obtain posterior probabilities used as the MRF data term. The MRF energy is minimized by α - β swap graph cut for obtaining labeling results.

Figure 4.23 shows labeling results. As we mentioned above, we used the half of dataset (31 images) are used for training. The total number of nodes for training is 747,937 and the primal solver for SVM training is necessary. The numbers of nodes of each test images are also shown

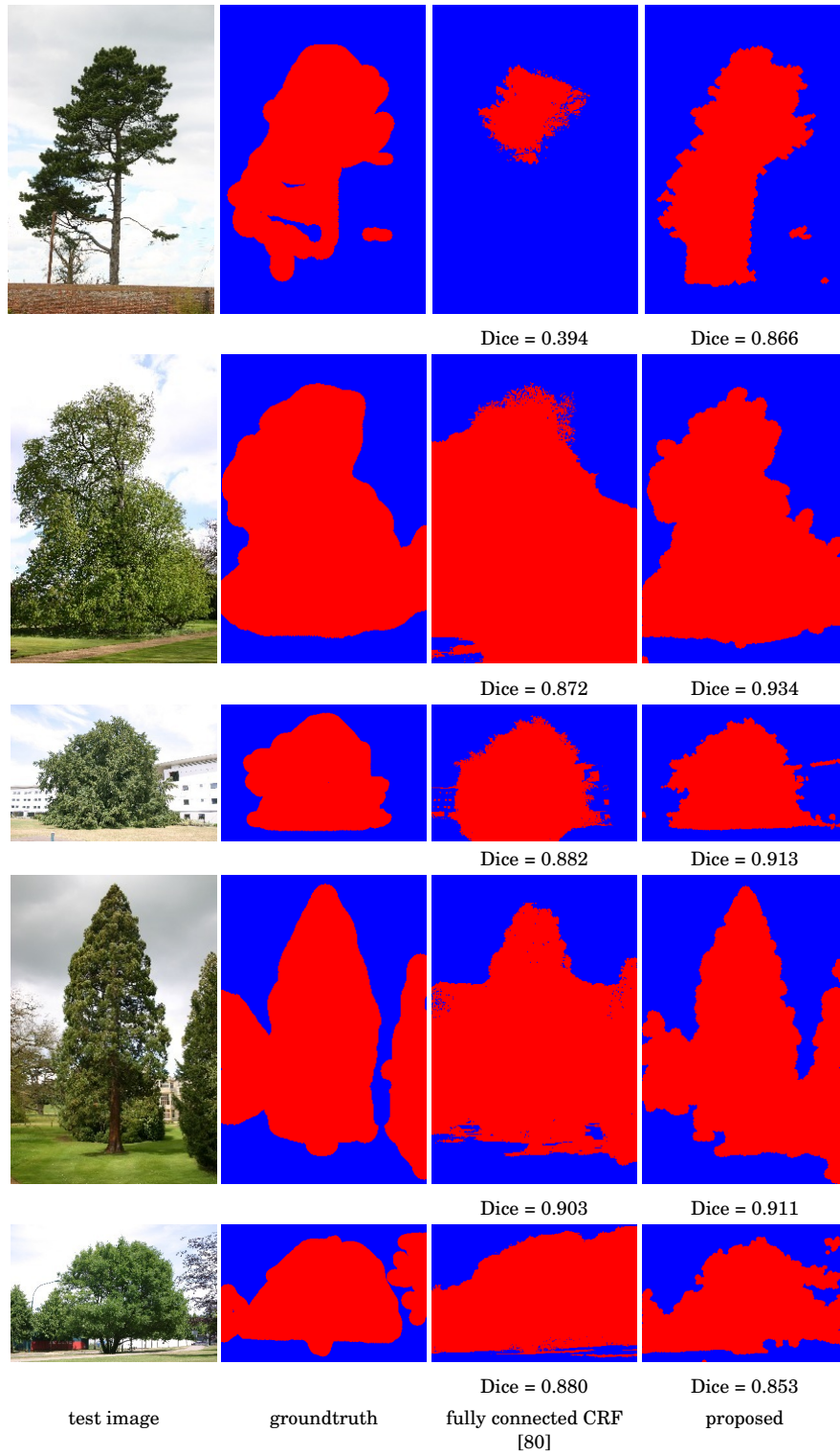


Figure 4.19: Labeling results on a subset 2 of the MSRC-21 dataset. Dice coefficient is shown below the images.

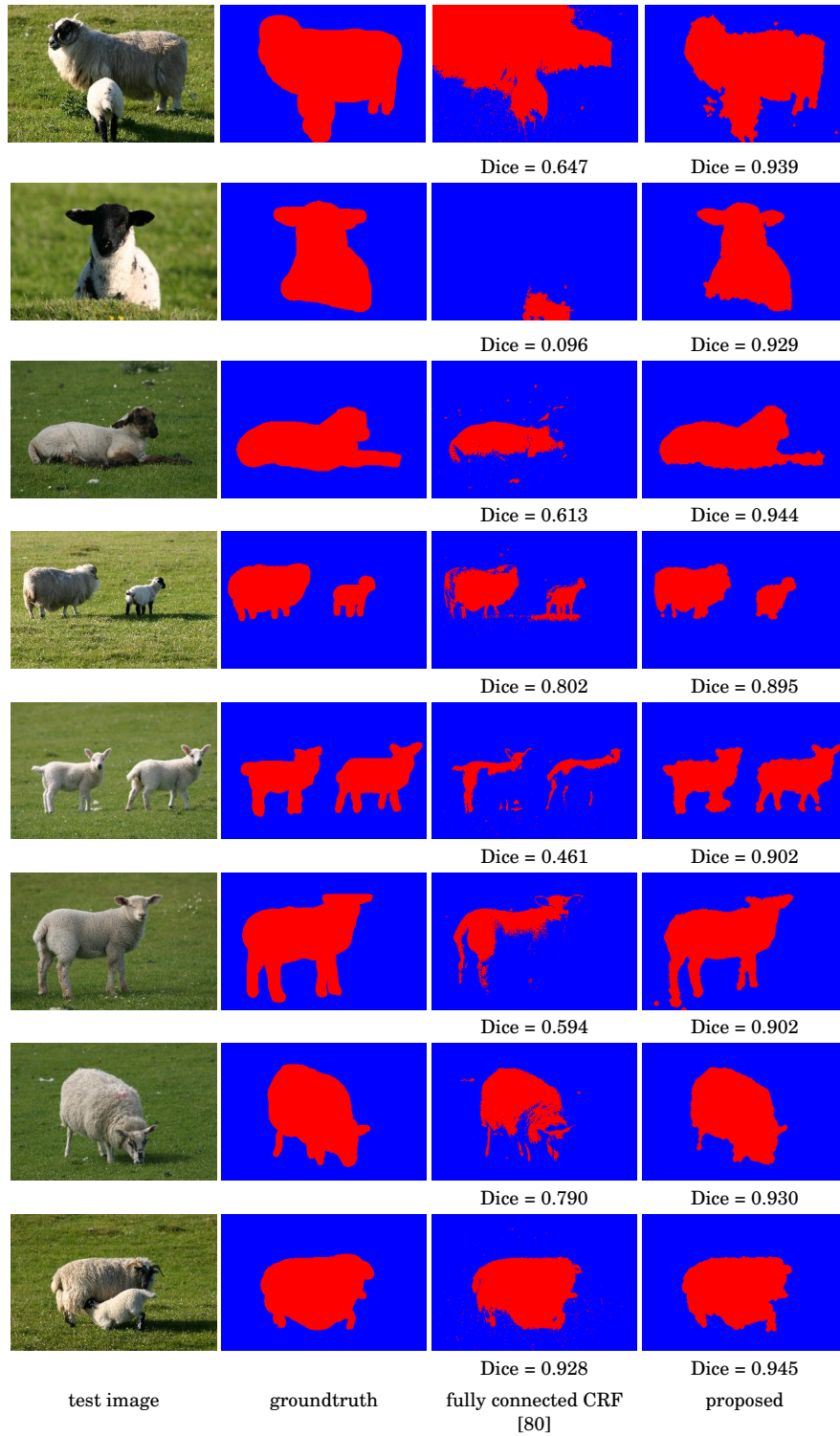


Figure 4.20: Labeling results on a subset 9 of the MSRC-21 dataset. Dice coefficient is shown below the images.

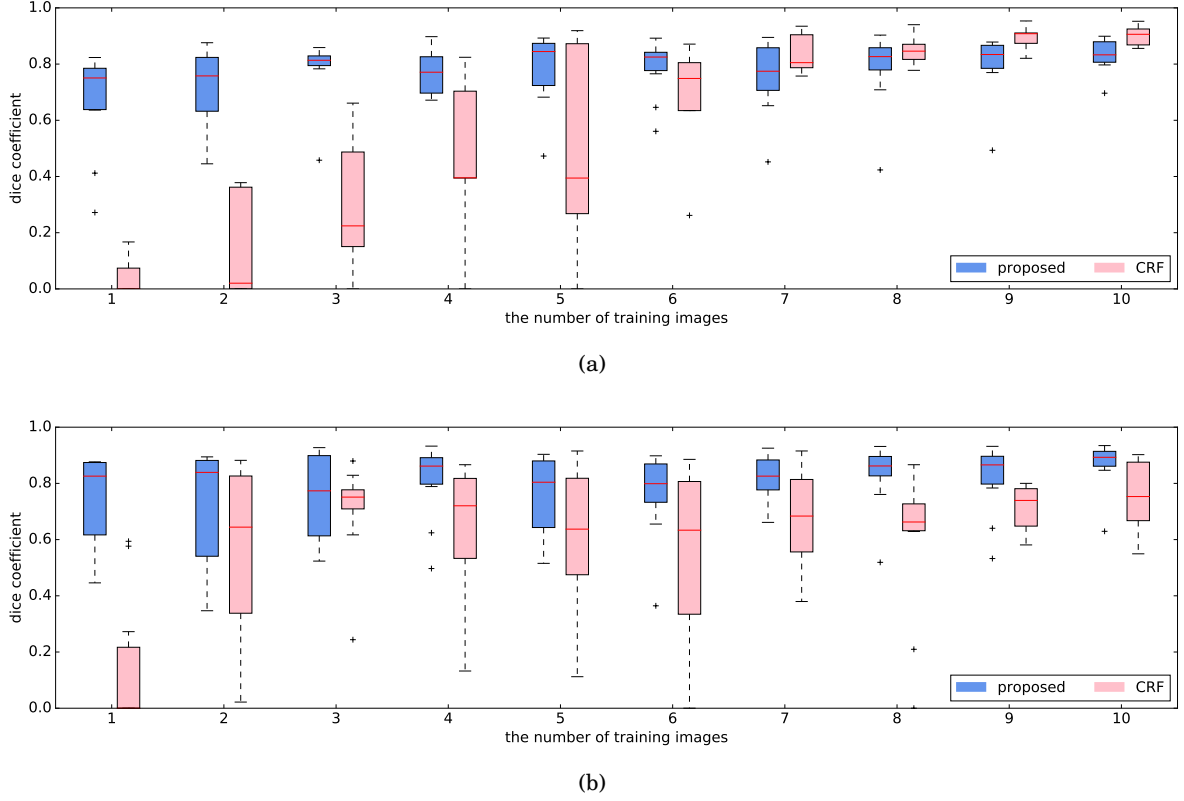


Figure 4.21: Box plots for Dice coefficients over different numbers of training images on (a) subset 2 and (b) subset 9 of the MSRC-21 dataset. The horizontal and vertical axes show the number of training images and the Dice coefficients, respectively.

in Figure 4.23. In SVM-MRF segmentation, labeling results are poor because the accuracy of MRF-based approaches highly depends on the data term. In other words, failures by the SVM classifier have the large impact on the poor accuracy. The conventional method provides cluttered labeling results because it classifies even small nodes. For instance, in the first two rows shows that the results of the conventional method provide small foreground regions. Meanwhile the proposed method can suppress the cluttered labels by selecting discriminative subtrees. In the middle and last two rows, foreground shapes of the proposed results are similar to the ground truth.

For quantitative evaluation, we used the dice coefficient [25]. Table 4.1 shows dice coefficients of each method. For conventional and proposed methods, we tested the procedures mentioned above repeatedly ten times and for the SVM-MRF method we tested only once. Note that the dice coefficient is calculated only for samples containing foreground. We can see that the proposed method outperforms the other two methods because using discriminative subtrees suppresses cluttered labels.

The proposed method outperforms the others in both the qualitative and quantitative evaluations. However, we need to discuss failure labeling results. Some failure examples are shown in

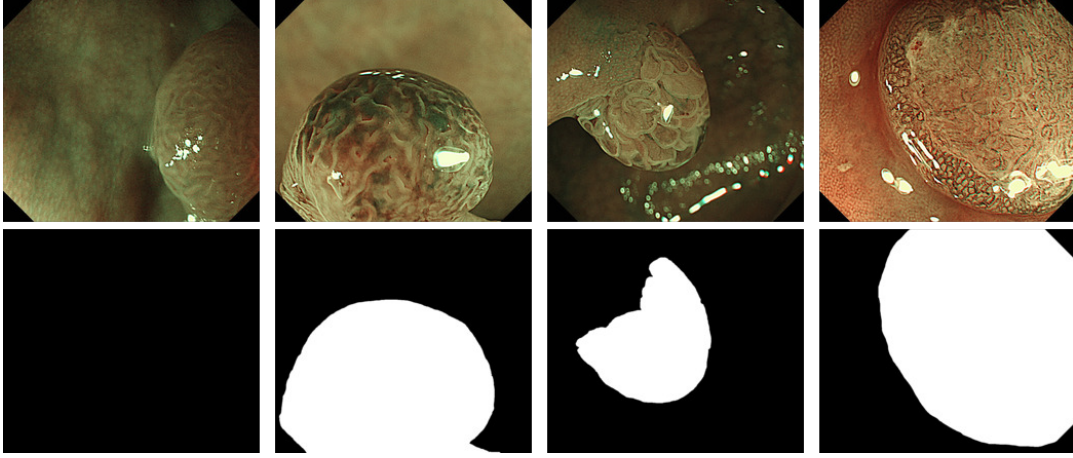


Figure 4.22: Examples of images in the NBI endoscopic image dataset. Upper row shows NBI images and bottom row shows corresponding masks. White color of the mask represents foreground and black represents background. The left-most image is a negative sample which doesn't have any foreground region.

Table 4.1: Dice coefficients of labeling results on NBI endoscopic images.

Method	dice coefficient
SVM-MRF	0.555
conventional	0.522 ± 0.056
proposed	0.653 ± 0.046

Figure 4.24. In the case of top row, the image is almost labeled as foreground. A possible reason is that the used histogram is simply constructed from four low level features, which might be too few to be discriminative enough. Therefore, using richer texture features is included in our future work. The proposed method labels as background inside of the foreground region in the middle row. In our method, subtrees are selected by one threshold, but optimal thresholds may be different for different images, which is a limitation of the proposed method. Results of bottom row provide small foreground labels, which correspond to specular reflections (highlight) and the surrounding regions. Because highlights are large area nodes, texture features extracted from highlights may affect classification results, and dealing with highlights is also one of our future work.

4.5 Summary

In this chapter, we proposed a labeling method for texture image segmentation that works with a few training images. Our method is based on a tree of shapes and histogram features derived from the tree structure and selects optimal discriminative subtrees for tree node classification. This is formulated as a joint optimization problem for estimating the threshold and classifier

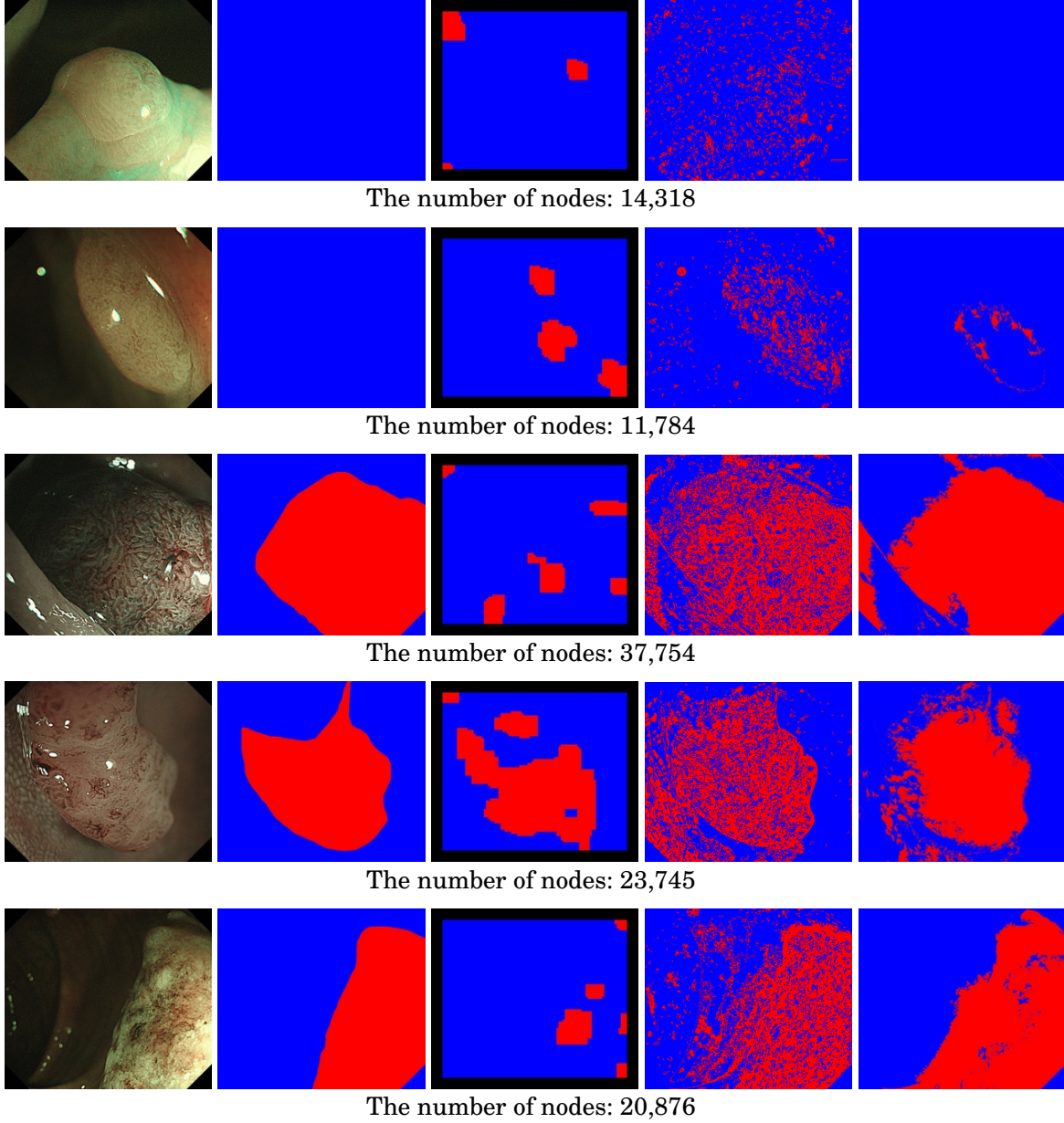


Figure 4.23: Labeling results. From left to right: test image, ground truth, labeling result of SVM-MRF, conventional, and proposed. The number of nodes in the trees of shapes created from test images are shown below the images. Red color represents foreground and blue background. Black color of SVM-MRF results represents unlabeled region due to the boundary effect.

parameters and is solved using iterative block-coordinate decent. Then, images are labeled using the estimated parameters and the tree of shapes by classifying each node from the root node to leaf nodes and then mapping classification results into the corresponding pixels. We evaluated the proposed method on the three datasets: a synthetic texture image datasets based on the UIUC database, the MSRC-21 dataset, and NBI endoscopic images. Experimental results show that

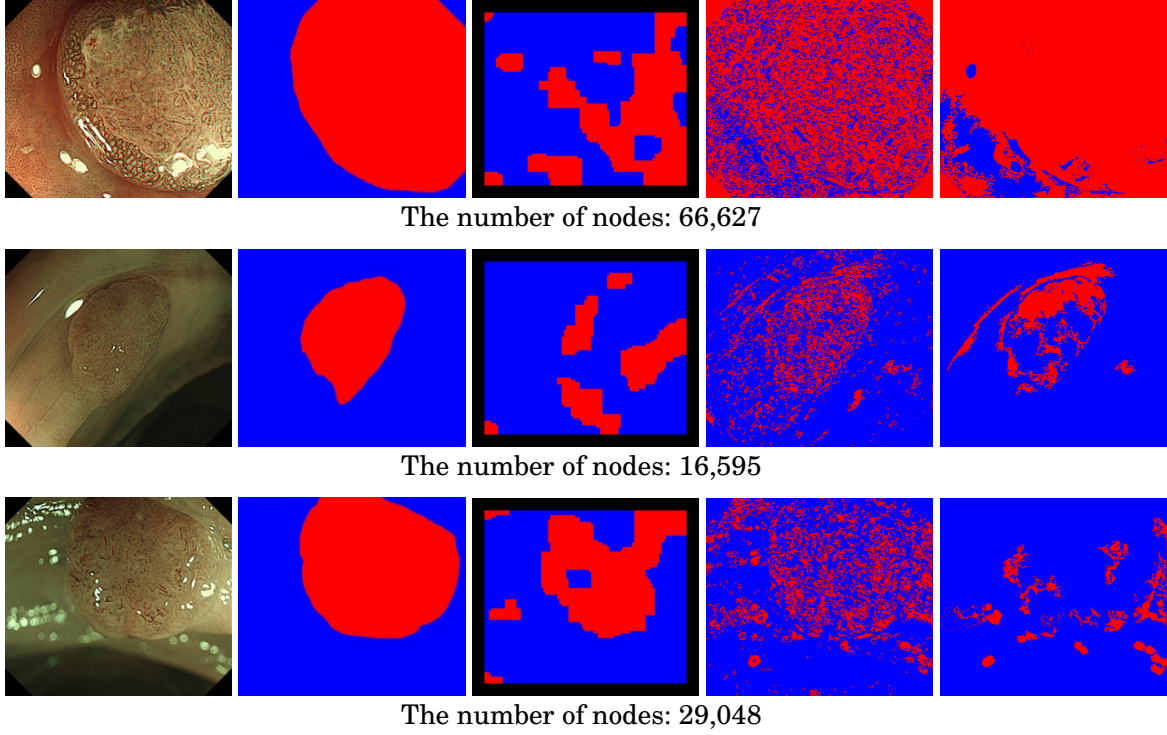


Figure 4.24: Some failure examples. From left to right: test image, ground truth, labeling result of SVM-MRF, conventional, and proposed. The number of nodes in a tree of shapes created from the test image is shown in the bottom of images. Red color represents foreground and blue background. Black color of SVM-MRF results represents unlabeled region due to boundary effect.

the proposed method outperforms other methods and provides more reliable results. Our future work includes improving the sample weights, extending our method to a multi-class problem, and seeking a more effective form of the labeling procedure using the hierarchical structure.

CONCLUSION

Colorectal endoscopy is widely used throughout world to detect colorectal cancer. Intra/inter-observer variability shows that visual inspection using NBI magnification findings can be subjective and require endoscopist's experience and hence a computalized system that provides an objective measure to endoscopists would be greatly help to endoscopists during examination. To this end, we developed an NBI videoendoscopy recognition system as our prior work. This system classifies the center of endoscopic video frame and display classification results on a monitor in a frame by frame manner. However, this system has two critical problems. The first problem is instability of classification results. Even if the video sequence continues to capture the same tumor, the classification results are highly unstable, that would be difficult to the status of tumors. The second problem lies in the fact that they can only a part of images of the video frame. In case that there a tumor is not in the center of the frame or multiple tumors exist in the frame, the system cannot provide appropriate objective measures.

In this thesis, we have developed three methods to improve problems of an NBI videoendoscope recognition system. The first method attempted the instability of classification results and we proposed D-DPF, a temporal smoothing method of posterior probabilities based on a particle filter with Dirichlet distribution. We introduced defocus information of a video frame as a confidence of a defocused frame. Experimental results with NBI endoscopic videos show that D-DPF can suppress instability of posterior probabilities.

The second method is an SVM-MRF image labeling method to recognize a whole endoscopic video frames (or images). This method used a posterior probabilities obtained from an SVM trained with NBI patches as a data term of MRF model. Moreover, highlight regions in an endoscopic video frames are considered to improve labeling result. Experimental results shows that labeling results become better by considering highlight regions, but a further improvement

is necessary.

In the third method, we tackled the image labeling problem again. We proposed a novel image labeling problem on the basis of a tree of shapes and histogram features computed on the tree structure. In a tree of shapes, we select subtrees to be useful for image labeling and then image labeling is done by classifying the selected subtrees. To select such optimal subtrees and to train a classifier, we defined a joint optimization problem of thresholds to select subtrees and parameters of an SVM classifier. Those parameters are estimated by a block-coordinate decent algorithm with respect to the thresholds and SVM parameters with training images. For quantitative evaluation, we used a synthetic texture image dataset and the MSRC-21 dataset and achieved a better labeling performance. An experiment with NBI endoscopic images shows that this method outperforms the SVM-MRF method.

In our future work, D-DPF in this thesis should be embedded in our NBI videoendoscopy recognition system, so that endoscopists can use these method during examinations. Embedding the two labeling methods into the system would might be further help for endoscopists. However, the computational cost of these methods are rather expensive to work in a real time, and displaying the labeling results frame by frame would be indistinct. Therefore, the two image labeling methods should be implemented as an application software that can be used for training of inexperienced endoscopists and for clinical experience.



DEVELOPMENT OF REAL-TIME CLASSIFICATION SYSTEM

A computer-aided diagnosis (CAD) system that provides an objective measure of the status of a tumor can be greatly help for endoscopists during examination. Therefore, we have developed NBI videoendoscopy recognition system as a prior work and proposed D-DPF in the body part of this thesis. However, these extensions have only been applied to offline videos (i.e., reading a stored movie file and processing each video frame) and have not yet been used or validated in actual clinical examinations.

This appendix chapter describes a newly developed CAD system that provides a real-time objective measure to endoscopists during examinations. Our system captures the online video stream from a videoendoscope via a video capture board-equipped desktop computer, converts the video format, and classifies each frame using a pretrained patch-based classifier [145]. Finally, the obtained classification results are displayed on a monitor. In the next section, we describe the CAD system design that enables the system to be mobile and compatible with different endoscopic systems in a hospital, the software part of the system, and the patch-based-classifier. We have specified the following three requirements that the developed CAD system must satisfy: mobility, high frame rate, and medical significance. Experimental results show that our system has a mobility and a sufficiently fast processing speed. In terms of medical significance, a requirement for the accuracy of the real-time assessment by endoscopists has been provided by a medical society [129], which is discussed in the Results and Discussion section. Following six months of clinical case studies of the developed CAD system in actual endoscopic examinations at the Hiroshima University Hospital, we showed that it is our system allows nonexpert endoscopists to diagnose with sufficient accuracy needed to meet the specified requirements.

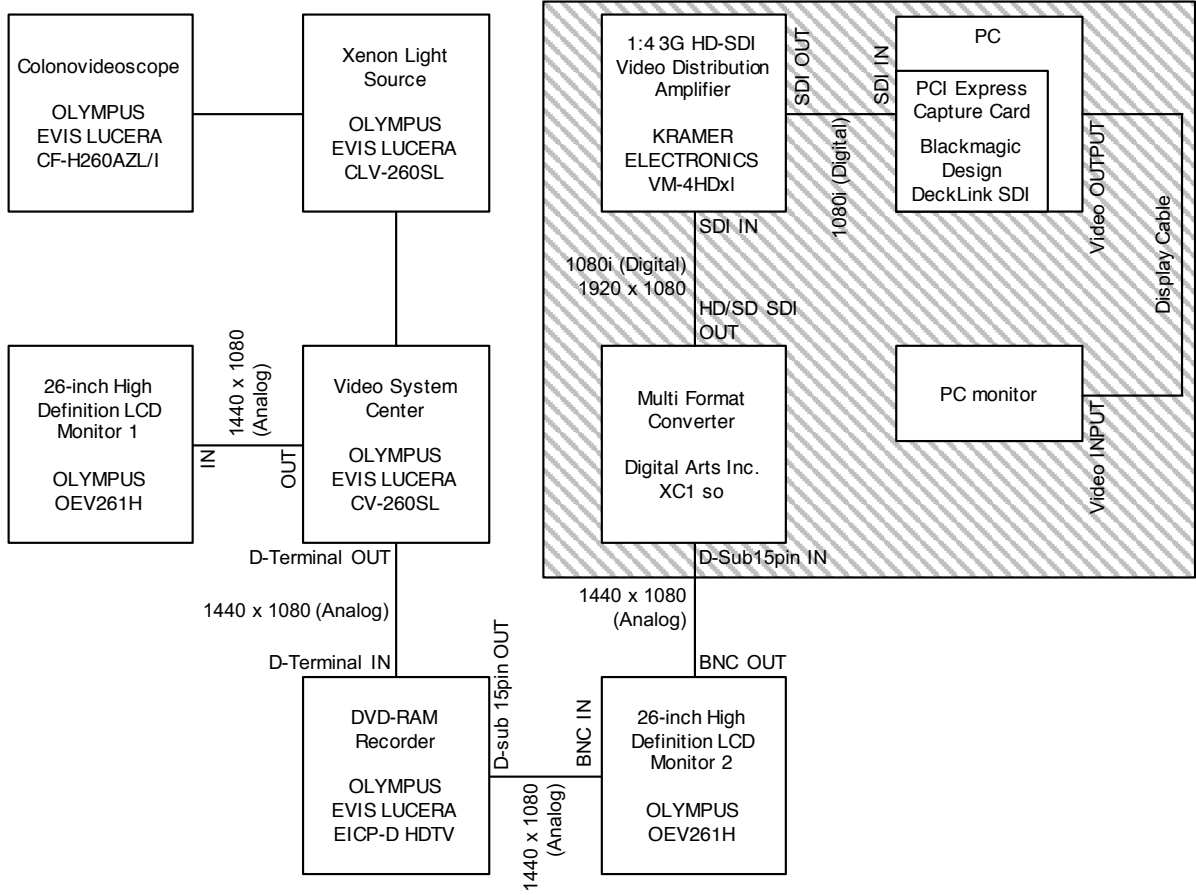


Figure A.1: An illustration of the system configuration of the endoscopic system (Olympus EVIS LUCERA) and our developed CAD system (gray shaded area).

A.1 Design of the Video Stream Capturing System

Herein, we describe the system design of the developed CAD system. There are two requirements the system needs to meet. First, we cannot modify the current configurations of the endoscopic systems that are regularly used in the hospital for our developed system and experiments. In this study, we used two different endoscopic systems: Olympus EVIS LUCERA (Figure A.1) and Olympus EVIS LUCERA ELITE (Figure A.2). These endoscopic systems have been configured and adjusted for regular examinations. Changing the configuration, e.g., switching the cables to intercept video streams, could cause the actual clinical flow between the endoscopy and the storage for medical data to stop. Second, the developed system must be able to deal with the two different endoscopic systems. In general, there are different endoscopic devices of different generations and types in different hospitals or even in a single hospital; hence, the developed system should have good mobility within the hospital. Keeping this in mind, we designed our system to capture the video stream from the videoendoscopes in such a way that simply attaching or detaching the connector for the video branch would be sufficient for operation.

A.1. DESIGN OF THE VIDEO STREAM CAPTURING SYSTEM

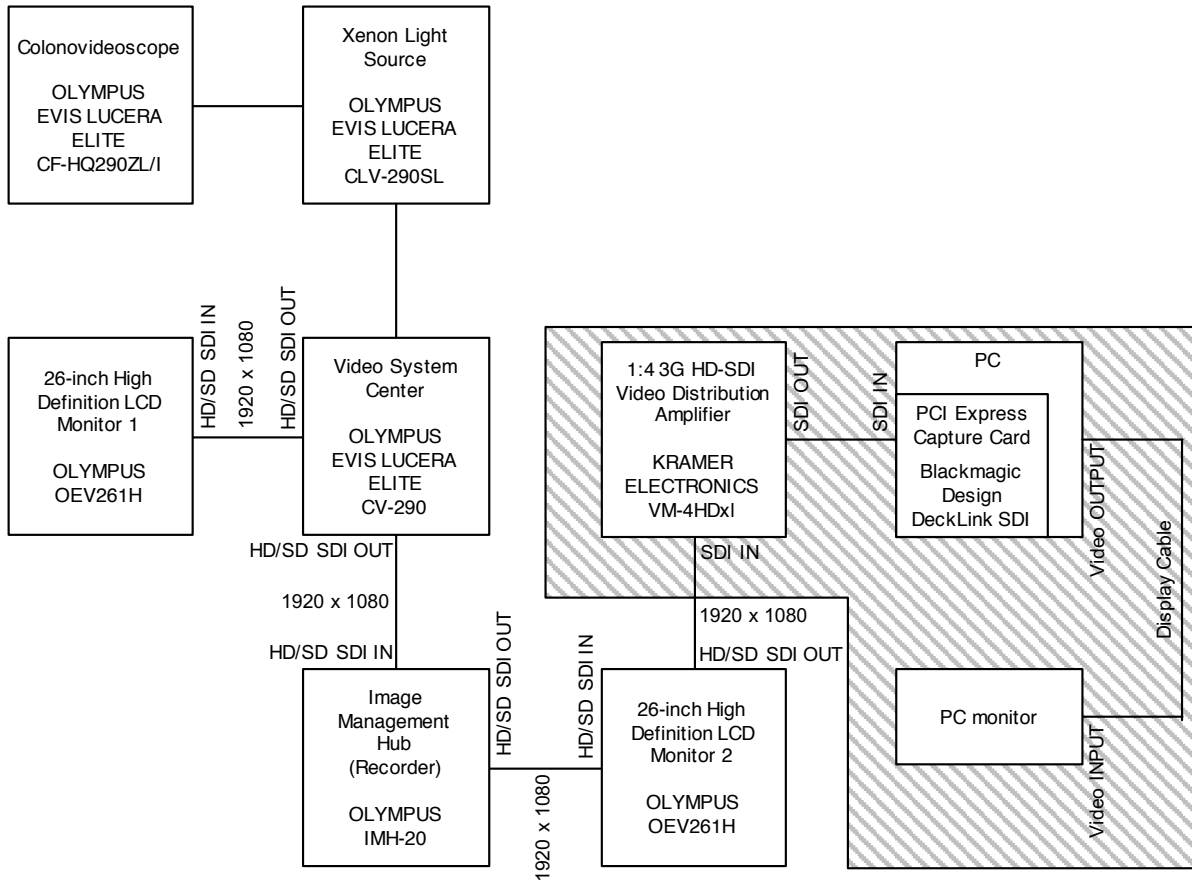


Figure A.2: An illustration of the system configuration of the endoscopic system (Olympus EVIS LUCERA ELITE) and our developed CAD system (gray shaded area).

The structure of the developed CAD system for one of the endoscope systems (Olympus Optical Co, Ltd; EVIS LUCERA) is shown in Figure A.1. The video stream captured by a colonovideoscope (Olympus Optical Co, Ltd; CF-H260AZL/I) with a xenon light source (Olympus Optical Co, Ltd; CLV-260SL) is sent to the video system center (Olympus Optical Co, Ltd; CV-260SL) to be processed. The processed stream is transferred to a digital versatile disc (DVD)- random access memory (RAM) recorder (Olympus Optical Co, Ltd; EICP-D HDTV) and is then passed to the second monitor (Olympus Optical Co, Ltd; OEV261H). The analog (RGB) video stream from the DVD-RAM recorder is displayed on the second monitor.

To capture the video stream for our system, we used the bypassed video stream from the second monitor. Because the video stream is analog in all the connections of this endoscopic system, we needed to convert it to a digital video stream using a multiformat video converter (XC1 co; Digital Arts Inc.); this conversion degrades the image quality. We then split the converted digital video stream by inserting a distribution amplifier (VM-4HDxI; Kramer Electronics Ltd.) and transferred it to a desktop computer equipped with a peripheral component interconnect (PCI) express video capture card (DeckLink SDI; Blackmagic Design Pty. Ltd.).

Figure A.2 shows the structure of the developed CAD system for another endoscope system (Olympus Optical Co, Ltd; EVIS LUCERA ELITE) comprising a colonovideoscope (Olympus Optical Co, Ltd; CF-HQ290 ZL/I), a xenon light source (Olympus Optical Co, Ltd; CLV-290SL), a video system center (Olympus Optical Co, Ltd; CV-290), a recorder (Olympus Optical Co, Ltd; IMH-20), and monitors (Olympus Optical Co, Ltd; OEV261H).

Similar to the first system, which is shown in Figure A.1, the video stream bypassed at the monitor is branched by the video distribution amplifier and is then transferred into the PCI express video capture card on the desktop computer. Note that a video converter is not required because the video stream is digital in all the connections of this endoscopic system.

At the end of the flow in the developed CAD system, the video stream is captured using the software development kit (SDK) of the capture card (DeckLink SDK 10.0; Blackmagic Design Pty. Ltd.). Then, the video frame is converted from YUV422 format to RGB format and is stored using an image-processing library (OpenCV 3.0 developer version; OpenCV.org) [12]. This color conversion is necessary because the digital video stream is usually represented in YUV color space whereas the software usually uses RGB color space for image processing. After the color conversion, the converted RGB video frame is passed to the patch-based classifier to compute the results. This is described in the following subsection.

A.2 Implementation of an Endoscopic Video Frame Classification System

An overview of the online classification of the video frames is already shown in Figure 1.10. Hence, we introduce the implementation and development environment of the classification system here.

The region-of-interest (ROI) is set to a rectangular patch at the center of the frame of the videoendoscope, and then densely sampled SIFT descriptors are extracted in the ROI. The extracted SIFT descriptors are represented as a histogram of the visual words (representative SIFT features) computed by hierarchical k-means clustering [116]. Each bin of this histogram is linearly scaled with a fixed factor (determined at the training phase; see below) because the scaling is well-known to affect the classification performance. This histogram feature (usually called the BoVW histogram feature) is then classified by a pretrained linear SVM classifier [21, 134, 142, 153] to obtain the classification probabilities for each category. These results (probabilities) are finally displayed on a monitor by superimposing the results onto the captured video frame. In this implementation, we used VLFeat 0.9.18 [154] for extracting the SIFT descriptors and for hierarchical k-means clustering, and LIBSVM 2.91 [19] for the linear SVM classifier. This online classification was developed under an integrated development environment (IDE; Visual Studio 2012; Microsoft Corp.) on a desktop computer (Intel Core i7-4770 3.4 GHz CPU with 16 GB memory, Microsoft Windows 7 Home Premium SP1 64bit) and written in C++.

The classifier-training phase was conducted offline before the online classification. We trained

a linear SVM classifier using the training samples mentioned above, and all the necessary parameters of the SVM classifier were stored in a file. During the online classification phase, this file was loaded and used for the SVM classification. This training phase was implemented on a different desktop computer (Intel Xeon CPU E5-2620 with 128 GB memory, Ubuntu 14.04 LTS; Canonical Ltd.) with several different codes written in C++ using the same libraries as the ones used in online classification. These codes were integrated with Bash and Perl 5.18 scripts for parallel processing.

Possible problem that might arise if our system is continuously used for several years is the discrepancy between the training samples and test samples due to a difference in the endoscopic devices. University hospitals are likely to replace endoscopic devices after a specific period of time. In that case, we need to collect as many training samples as possible for the new device to obtain acceptable classification results by training classifiers with the newly collected samples. However, it is impractical to collect a large number of training samples in a short period of time for each endoscope. To overcome this problem, we use a transfer learning-based learning approach proposed by Sonoyama et al. [139, 140] that trains a classifier with samples of the new endoscope by reusing (or transferring) the training samples from the old one. This enables us to maintain the classification performance without recollecting training samples when switching our system to a new endoscope.

A.3 Results and Discussions

A snapshot of the endoscopic system and the developed CAD system is shown in Figure A.3. The developed system consists of a desktop computer, a monitor, a keyboard, and a mouse arranged in a single mobile rack. The use of the PCI video capture card makes the system look slightly large; however, it is not difficult to move the rack from one endoscope to another. To make it smaller, we could develop a system on a laptop computer using a small video-capture device (e.g., the UltraStudio Mini Recorder for Thunderbolt; Blackmagic Design Pty. Ltd.). In addition, we are currently developing a hardware implementation of the online classification and displaying system on a field-programmable gate array (FPGA), which is an integrated circuit designed to be configured by a customer or a designer after manufacturing [77]. This will allow the manufacturing of a pocket-sized device that receives digital video signals, classifies video frames, and directly outputs the results to the monitor without using desktop or laptop computers.

Figure A.4 shows sample screen shots of a monitor displaying the results of the developed CAD system. The size of the captured video frame in which the endoscopic video stream is displayed is full HD (1980×1080 pixels). In each video frame, an ROI patch of 200×200 pixels (shown as white squares in Figure A.4) at the center of the endoscopic video stream is trimmed and classified by the pretrained SVM classifier. The classification result is shown on the left side of the endoscopic video frame. In Figure A.4(a), the result of the three-category classification for this video frame is shown as “type B,” which is the category label given by the classifier, followed by the three probabilities of each category: 1.1% for type A, 98.9% for type B, and no probability for type C. Therefore, our system provides an objective measure indicating that this frame is type

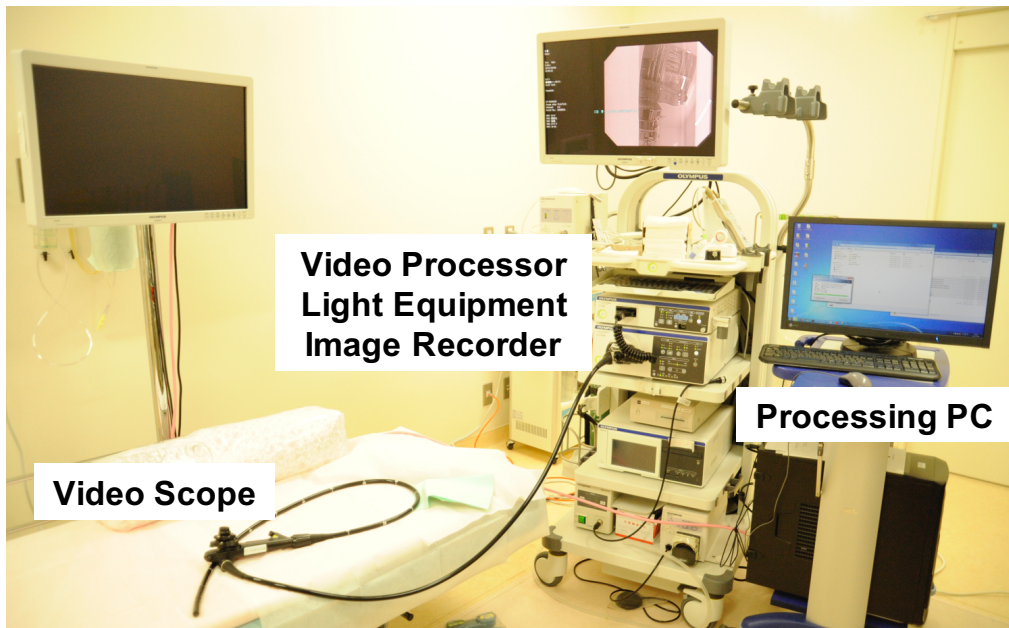


Figure A.3: The endoscopic system and the developed system.

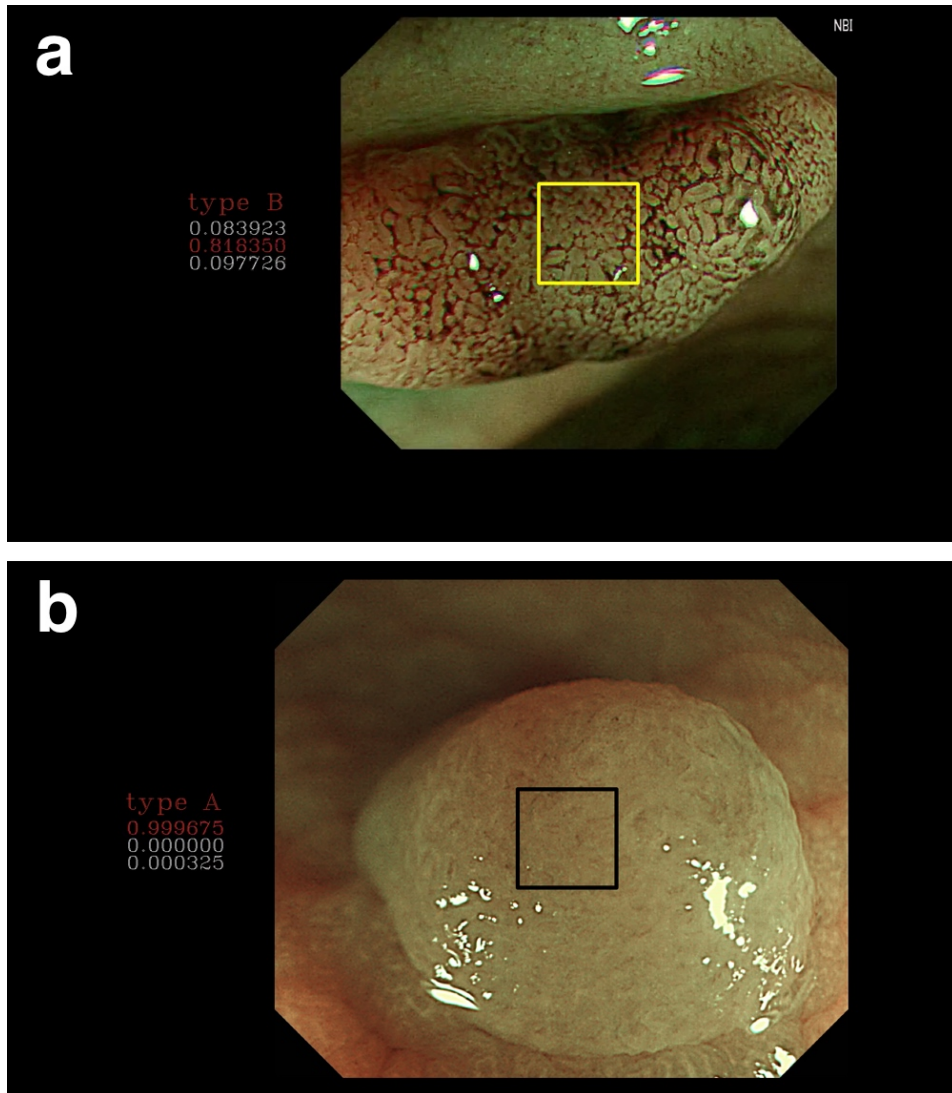


Figure A.4: Screen shots of the developed video frame classification system. The white square is the ROI to be classified. The results are superimposed on a video frame of (a) the endoscopic system EVIS LUCERA (shown in Figure 2) and (b) the endoscopic system EVIS LUCERA ELITE (shown in Figure 3). The frame size in both cases is 1920×1080 pixels. The sizes of the endoscopic video stream are (a) 1000×870 pixels and (b) 1156×1006 pixels. Note that the black background area is used to show information from the endoscopic systems such as the date, time, ID numbers, and video frame snapshots, which have not been shown here because of confidentiality reasons.

B with 98% confidence, while it may be type A with probability of 1.1%. In Figure A.4(b), the result of the two-category classification (type A or not) is shown as “type B” (which means “not type A”) with a probability of 98.8%, and there is still a probability of 1.2% that this frame is type A.

The current system processes frame by frame; therefore, the probability results displayed on the monitor could become too unstable for endoscopists to determine the system output. This

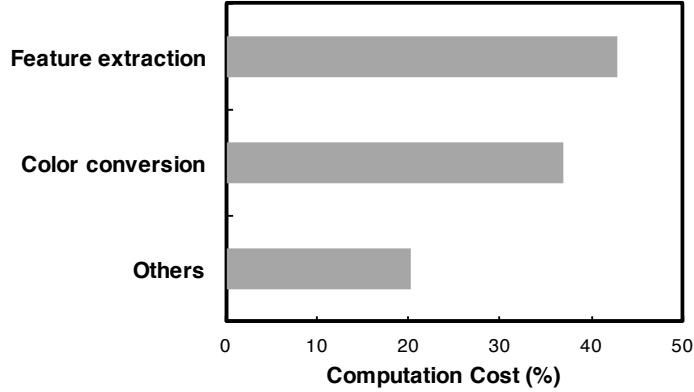


Figure A.5: Computational cost per frame.

could be because of the incorrect classification results resulting from motion blur, out of focus images, or color bleeding. To make the classification results stable, we proposed the following two methods: a stable labeling method that could suppress frequent label changes [61] and D-DPF. These methods will be included in our system in future.

The throughput of the developed system and the endoscopic video stream are approximately about 20 fps (approximately 50 ms per frame) [78] and 30 fps, respectively. Figure A.5 shows the relative computational cost per frame. We can see that the SIFT feature extraction and the color conversion from YUV422 to RGB require the bulk of the computation time. Further optimization of these processes needs to be done to achieve a better system throughput of up to 30fps. This is also our future work. However, the current throughput of 20 fps is high enough for normal clinical use. Note that, there is a tradeoff between the ROI rectangle size, classification performance, and processing time. If the ROI size is too small, the classification performance would decrease because of the insufficient number of SIFT features being extracted. However, classifying ROIs of larger sizes makes the classification results reliable even though the processing time increases. The current ROI size of 200×200 pixels is an acceptable compromise for the current result.

Apart from the aforementioned engineering aspect of the developed CAD system described above, the medical significance of our system is also important. Herein, we refer to statements concerning the real-time endoscopic assessment of the histology of diminutive (≤ 5 mm) colorectal polyps published by the Preservation and Incorporation of Valuable Endoscopic Innovations (PIVI) committee of the American Society for Gastrointestinal Endoscopy [129]. One of the two PIVI recommendations states, “in order for a technology to be used to guide the decision to leave suspected rectosigmoid hyperplastic polyps ≤ 5 mm in size in place (without resection), the technology should provide $\geq 90\%$ negative predictive value ... for adenomatous histology” [129]. In other words, in the context of our system, more than 90% of lesions diagnosed as “type A” by community endoscopists using our CAD system should histologically be non- neoplastic lesions. On the basis of this recommendation, our developed system was introduced in actual

endoscopic examinations in the hospital and was evaluated for its medical significance. Details can be found in Kominami et al. [78]; however, we highlight their results here. These clinical case studies were conducted for six months (between October 2014 and March 2015). Endoscopists used our developed system to classify video frames of colon tumors into two categories, i.e., type A (non-neoplastic lesions) or types B and C3 (neoplastic lesions). Classification probabilities obtained from our developed system were evaluated in concordance with the two diagnostic results: endoscopic and histological diagnoses. For the concordance with the endoscopic diagnosis, the concordance rate was 96.6% with a kappa static value of 0.93 and a 95% confidence interval of 0.89–1.00. For the concordance with histology, Kominami et al. performed the Mann-Whitney U test and obtained an accuracy of 93.2% (sensitivity: 93.0%, specificity: 93.3%, positive predictive value: 93.0%, and negative predictive value: 93.3%). Thus, these results show that nonexpert endoscopists were able to diagnose colon tumors with an accuracy sufficient to satisfy the PIVI requirement using our system.

While the endoscopists concluded that our CAD system may satisfy the PIVI recommendations, they also said that “further development of our real-time image recognition system ... and additional studies aimed at assessing whether community endoscopists may successfully meet both PIVI thresholds are needed” [78]. The motivation to publish the current paper on the system development arises from the necessity to further develop the system for clinical studies. It has been reported [86, 125, 128] that the performance of endoscopists is sufficient to achieve the PIVI requirement but only with a prior training module. These reports highlight the importance of the studies on the performance of non-expert endoscopists in actual clinical examinations [125], and on the necessity of training for maintaining the endoscopist’s performance [86, 125]. We believe that the our CAD system will be useful in future clinical studies and in training and assessing the skills of endoscopists.

A.4 Summary

We developed a real-time colorectal tumor classification system that provides a real-time objective measure of the status of colon tumors to endoscopists during examinations. This system was built in such a way that no modifications to the actual endoscopic systems being used in hospitals are necessary. A six-month-long clinical case study using the developed system for actual endoscopic examinations demonstrated that our system is mobile in the hospital, a processing speed of 20 fps is sufficient for examinations, and the system is medically significant from the viewpoint of the PIVI recommendations. Our future work will include making the system faster, more compact, and more user-friendly so that the system could be used by community endoscopists.

BIBLIOGRAPHY

- [1] M. ARNOLD, A. GHOSH, G. LACEY, S. PATCHETT, AND H. MULCAHY, *Indistinct Frame Detection in Colonoscopy Videos*, in Machine Vision and Image Processing Conference, 2009. IMVIP '09. 13th International, Sept 2009, pp. 47–52.
- [2] S. BANK, J. S. COBB, D. G. BURNS, AND I. N. MARKS, *Dissecting microscopy of rectal mucosa*, The Lancet, 295 (1970), pp. 64–65.
- [3] G. L. BEETS AND R. G. H. BEETS-TAN, *Pretherapy imaging of rectal cancers: ERUS or MRI?*, Surg Oncol Clin N Am, 19 (2010), pp. 733–741.
- [4] M. B. BEHRENS, S. GROSS, AND A., *Chan-Vese-Segmentation of Polyps in Colonoscopic Image Data*, in Proceedings of the 15th International Student Conference on Electrical Engineering POSTER 2011, Prague, Czech Republic, May 12 2011.
- [5] J. BERNAL, F. J. SÁNCHEZ, AND F. VILARIÑO, *A Region Segmentation Method for Colonoscopy Images Using a Model of Polyp Appearance*, in IbPRIA, 2011, pp. 134–142.
- [6] L. BERTELLI, T. YU, D. VU, AND B. GOKTURK, *Kernelized structural svm learning for supervised object segmentation*, in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, June 2011, pp. 2153–2160.
- [7] C. M. BISHOP, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- [8] M. BISWAS AND D. DEY, *Bi-dimensional Statistical Empirical Mode Decomposition-Based Video Analysis for Detecting Colon Polyps Using Composite Similarity Measure*, in Intelligent Computing, Communication and Devices, L. C. Jain, S. Patnaik, and N. Ichalkaranje, eds., vol. 309 of Advances in Intelligent Systems and Computing, Springer India, 2015, pp. 297–308.
- [9] E. BORENSTEIN, E. SHARON, AND S. ULLMAN, *Combining top-down and bottom-up segmentation*, in 2004 Conference on Computer Vision and Pattern Recognition Workshop, June 2004, pp. 46–46.

BIBLIOGRAPHY

- [10] A. C. BOVIK, M. CLARK, AND W. S. GEISLER, *Multichannel texture analysis using localized spatial filters*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 12 (1990), pp. 55–73.
- [11] Y. BOYKOV, O. VEKSLER, AND R. ZABIH, *Fast approximate energy minimization via graph cuts*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23 (2001), pp. 1222–1239.
- [12] G. BRADSKI ET AL., *The opencv library*, Dr. Dobb’s Journal of Software Tools, (2000).
- [13] M. BREIER, S. GROSS, A. BEHRENS, T. STEHLE, AND T. AACH, *Active contours for localizing polyps in colonoscopic NBI image data*, in Proc. of Medical Imaging 2011: Computer-Aided Diagnosis, 2011, pp. 79632M–79632M–10.
- [14] J. V. CANDY, *Bayesian signal processing: classical, modern, and particle filtering methods*, Adaptive and learning systems for signal processing, communications, and control, Wiley, Hoboken, N.J., 2009.
- [15] CANER RESEARCH UK, *Bowel cancer statistics*, <http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer>, 2014, [Online; accessed 2017/03/02].
- [16] C. L. CANON, *Is there still a role for double-contrast barium enema examination?*, Clin Gastroenterol Hepatol, 6 (2008), pp. 389–392.
- [17] E. CARLINET AND T. GÉRAUD, *Mtos: A tree of shapes for multivariate images*, IEEE Transactions on Image Processing, 24 (2015), pp. 5330–5342.
- [18] C.-C. CHANG, C.-R. HSIEH, H.-Y. LOU, C.-L. FANG, C. TIONG, J.-J. WANG, I.-V. WEI, S.-C. WU, J.-N. CHEN, AND Y.-H. WANG, *Comparative study of conventional colonoscopy, magnifying chromoendoscopy, and magnifying narrow-band imaging systems in the differential diagnosis of small colonic polyps between trainee and experienced endoscopist*, Int J Colorectal Dis, 24 (2009), pp. 1413–9.
- [19] C.-C. CHANG AND C.-J. LIN, *Libsvm: A library for support vector machines*, ACM Trans. Intell. Syst. Technol., 2 (2011), pp. 27:1–27:27.
- [20] J. COUSTY AND L. NAJMAN, *Incremental Algorithm for Hierarchical Minimum Spanning Forests and Saliency of Watershed Cuts*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 272–283.
- [21] N. CRISTIANINI AND J. SHAWE-TAYLOR, *An introduction to support vector machines and other kernel-based learning methods*, Cambridge university press, 2000.

- [22] G. CSURKA, C. DANCE, L. FAN, J. WILLAMOWSKI, AND C. BRAY, *Visual categorization with bags of keypoints*, in Workshop on statistical learning in computer vision, ECCV, vol. 1, Prague, 2004, pp. 1–2.
- [23] C. D’ELIA, S. RUSCINO, M. ABBATE, B. AIAZZI, S. BARONTI, AND L. ALPARONE, *Sar image classification through information-theoretic textural features, mrf segmentation, and object-oriented learning vector quantization*, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7 (2014), pp. 1116–1126.
- [24] Y. DENG AND B. MANJUNATH, *Unsupervised segmentation of color-texture regions in images and video*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23 (2001), pp. 800–810.
- [25] L. R. DICE, *Measures of the amount of ecologic association between species*, Ecology, 26 (1945), pp. 297–302.
- [26] A. DUFOUR, O. TANKYEVYCH, B. NAEGEL, H. TALBOT, C. RONSE, J. BARUTHIO, P. DOK-LÁDAL, AND N. PASSAT, *Filtering and segmentation of 3d angiographic data: Advances based on mathematical morphology*, Medical Image Analysis, 17 (2013), pp. 147 – 164.
- [27] M. EVANS, N. HASTINGS, AND B. PEACOCK, *Statistical Distributions*, Wiley Series in Probability and Statistics, Wiley, 2000.
- [28] M. EVERINGHAM, L. VAN GOOL, C. K. I. WILLIAMS, J. WINN, AND A. ZISSERMAN, *The pascal visual object classes (voc) challenge*, International Journal of Computer Vision, 88 (2010), pp. 303–338.
- [29] R.-E. FAN, K.-W. CHANG, C.-J. HSIEH, X.-R. WANG, AND C.-J. LIN, *LIBLINEAR: A library for large linear classification*, Journal of Machine Learning Research, 9 (2008), pp. 1871–1874.
- [30] C. FARABET, C. COUPRIE, L. NAJMAN, AND Y. LECUN, *Learning hierarchical features for scene labeling*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (2013), pp. 1915–1929.
- [31] L. FEI-FEI AND P. PERONA, *A bayesian hierarchical model for learning natural scene categories*, in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), vol. 2, June 2005, pp. 524–531 vol. 2.
- [32] P. F. FELZENSZWALB AND D. P. HUTTENLOCHER, *Efficient graph-based image segmentation*, International Journal of Computer Vision, 59 (2004), pp. 167–181.
- [33] B. FULKERSON, A. VEDALDI, AND S. SOATTO, *Localizing Objects with Smart Dictionaries*, in Computer Vision – ECCV 2008, D. Forsyth, P. Torr, and A. Zisserman, eds., vol. 5302 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2008, pp. 179–192.

- [34] B. FULKERSON, A. VEDALDI, AND S. SOATTO, *Class segmentation and object localization with superpixel neighborhoods*, in 2009 IEEE 12th International Conference on Computer Vision, Sept 2009, pp. 670–677.
- [35] T. GÉRAUD, E. CARLINET, S. CROZET, AND L. NAJMAN, *A Quasi-linear Algorithm to Compute the Tree of Shapes of nD Images*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 98–110.
- [36] R. GIRSHICK, J. DONAHUE, T. DARRELL, AND J. MALIK, *Rich feature hierarchies for accurate object detection and semantic segmentation*, in 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 2014, pp. 580–587.
- [37] K. GONO, T. OBI, M. YAMAGUCHI, N. OHYAMA, H. MACHIDA, Y. SANO, S. YOSHIDA, Y. HAMAMOTO, AND T. ENDO, *Appearance of enhanced tissue features in narrow-band endoscopic imaging*, J Biomed Opt, 9 (2004), pp. 568–577.
- [38] K. GONO, K. YAMAZAKI, N. DOGUCHI, T. NONAMI, T. OBI, M. YAMAGUCHI, N. OHYAMA, H. MACHIDA, Y. SANO, S. YOSHIDA, Y. HAMAMOTO, AND T. ENDO, *Endoscopic Observation of Tissue by Narrowband Illumination*, Optical Review, 10 (2003), pp. 211–215.
- [39] N. GOPALSWAMY, S. NEWAZ, S. GIANTI, A. BHUTANI, R. J. MARKERT, AND L. SWAMY, *Digital rectal examination as a part of colorectal cancer screening in hospitalized veterans*, Am J Gastroenterol, 95 (2000), pp. 2534–2535.
- [40] N. GORDON, D. SALMOND, AND A. SMITH, *Novel approach to nonlinear/non-Gaussian Bayesian state estimation*, Radar and Signal Processing, IEE Proceedings F, 140 (1993), pp. 107–113.
- [41] S. GROSS, M. KENNEL, T. STEHLE, J. WULFF, J. TISCHENDORF, C. TRAUTWEIN, AND T. AACH, *Polyp Segmentation in NBI Colonoscopy*, in Bildverarbeitung für die Medizin 2009, H.-P. Meinzer, T. Deserno, H. Handels, and T. Tolxdorff, eds., Informatik aktuell, Springer Berlin Heidelberg, 2009, pp. 252–256.
- [42] S. GROSS, T. STEHLE, A. BEHRENS, R. AUER, T. AACH, R. WINOGRAD, C. TRAUTWEIN, AND J. TISCHENDORF, *A comparison of blood vessel features and local binary patterns for colorectal polyp classification*, in Proc. SPIE, vol. 7260, 2009, pp. 72602Q–72602Q–8.
- [43] GUNMA PREFECTURAL CANCER CENTER, *Joint investigation of survival rates in the member faculty of japan association of clinical cancer centers*, <http://www.gunma-cc.jp/sarukihan/seizonritu/seizonritu.html>, 2013, [Online; accessed 2017/03/02].

-
- [44] M. HÄFNER, A. GANGL, R. KWITT, A. UHL, A. VÉCSEI, AND F. WRBA, *Improving Pit-Pattern Classification of Endoscopy Images by a Combination of Experts*, in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009, G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor, eds., vol. 5761 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2009, pp. 247–254.
- [45] M. HÄFNER, A. GANGL, M. LIEDLGRUBER, A. UHL, A. VECSEI, AND F. WRBA, *Combining Gaussian Markov random fields with the discrete-wavelet transform for endoscopic image classification*, in Digital Signal Processing, 2009 16th International Conference on, July 2009, pp. 1–6.
- [46] M. HÄFNER, A. GANGL, M. LIEDLGRUBER, A. UHL, A. VÉCSEI, AND F. WRBA, *Pit pattern classification using extended local binary patterns*, in 2009 9th International Conference on Information Technology and Applications in Biomedicine, Nov 2009, pp. 1–4.
- [47] M. HÄFNER, A. GANGL, M. LIEDLGRUBER, A. UHL, A. VECSEI, AND F. WRBA, *Pit Pattern Classification Using Multichannel Features and Multiclassification*, IGI Global, Hershey, PA, USA, 2009, pp. 335–350.
- [48] M. HÄFNER, A. GANGL, M. LIEDLGRUBER, A. UHL, A. VÉCSEI, AND F. WRBA, *Classification of Endoscopic Images Using Delaunay Triangulation-Based Edge Features*, in Image Analysis and Recognition, A. Campilho and M. Kamel, eds., vol. 6112 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2010, pp. 131–140.
- [49] M. HÄFNER, A. GANGL, M. LIEDLGRUBER, A. UHL, A. VECSEI, AND F. WRBA, *Endoscopic Image Classification Using Edge-Based Features*, in Pattern Recognition (ICPR), 2010 20th International Conference on, Aug 2010, pp. 2724–2727.
- [50] M. HÄFNER, C. KENDLBACHER, W. MANN, W. TA FERL, F. WRBA, A. GANGL, A. VECSEI, AND A. UHL, *Pit Pattern Classification of Zoom-Endoscopic Colon Images using Histogram Techniques*, in Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic, June 2006, pp. 58–61.
- [51] M. HÄFNER, R. KWITT, A. UHL, A. GANGL, F. WRBA, AND A. VÉCSEI, *Feature extraction from multi-directional multi-resolution image transformations for the classification of zoom-endoscopy images*, Pattern Analysis and Applications, 12 (2009), pp. 407–413.
- [52] M. HÄFNER, R. KWITT, A. UHL, F. WRBA, A. GANGL, AND A. VÉCSEI, *Computer-assisted pit-pattern classification in different wavelet domains for supporting dignity assessment of colonic polyps*, Pattern Recognition, 42 (2009), pp. 1180–1191.
- [53] M. HÄFNER, R. KWITT, F. WRBA, A. GANGL, A. VECSEI, AND A. UHL, *One-against-one classification for zoom-endoscopy images*, in Advances in Medical, Signal and

- Information Processing, 2008. MEDSIP 2008. 4th IET International Conference on, July 2008, pp. 1–4.
- [54] S. HALLIGAN AND S. A. TAYLOR, *CT colonography: Results and limitations*, European Journal of Radiology, 61 (2007), pp. 400–408.
- [55] C. HE, T. ZHUO, X. SU, F. TU, AND D. CHEN, *Local topographic shape patterns for texture description*, IEEE Signal Processing Letters, 22 (2015), pp. 871–875.
- [56] HEALTH STATISTICS AND INFORMATICS DEPARTMENT, WORLD HEALTH ORGANIZATION, *Global burden of disease*, http://www.who.int/topics/global_burden_of_disease/en/, 2008, [Online; accessed 2017/03/02].
- [57] S. J. HEITMAN, F. AU, B. J. MANNS, S. E. MCGREGOR, AND R. J. HILSDEN, *Nonmedical Costs of Colorectal Cancer Screening With the Fecal Occult Blood Test and Colonoscopy*, Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association, 6 (2008), pp. 912–917.
- [58] K. HELD, E. R. KOPS, B. J. KRAUSE, W. M. WELLS, R. KIKINIS, AND H. W. MULLER-GARTNER, *Markov random field segmentation of brain mr images*, IEEE Transactions on Medical Imaging, 16 (1997), pp. 878–886.
- [59] N. HERVÉ, N. BOUJEMAA, AND M. E. HOULE, *Document description: what works for images should also work for text?*, in Proc. SPIE, vol. 7255, 2009, pp. 72550B–72550B–12.
- [60] R. HIGASHI, T. URAOKA, J. KATO, K. KUWAKI, S. ISHIKAWA, Y. SAITO, T. MATSUDA, H. IKEMATSU, Y. SANO, S. SUZUKI, Y. MURAKAMI, AND K. YAMAMOTO, *Diagnostic accuracy of narrow-band imaging and pit pattern analysis significantly improved for less-experienced endoscopists after an expanded training program*, Gastrointest Endosc, 72 (2010), pp. 127–35.
- [61] T. HIRAKAWA, T. TAMAKI, B. RAYTCHEV, K. KANEDA, T. KOIDE, S. YOSHIDA, Y. KOMINAMI, T. MATSUO, R. MIYAKI, AND S. TANAKA, *Labeling colorectal nbi zoom-videoendoscope image sequences with mrf and svm*, in 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), July 2013, pp. 4831–4834.
- [62] M. HIRATA, S. TANAKA, S. OKA, I. KANEKO, S. YOSHIDA, M. YOSHIHARA, AND K. CHAYAMA, *Evaluation of microvessels in colorectal tumors by narrow band imaging magnification*, Gastrointest Endosc, 66 (2007), pp. 945–52.

- [63] M. HIRATA, S. TANAKA, S. OKA, I. KANEKO, S. YOSHIDA, M. YOSHIHARA, AND K. CHAYAMA, *Magnifying endoscopy with narrow band imaging for diagnosis of colorectal tumors*, *Gastrointestinal Endoscopy*, 65 (2007), pp. 988 – 995.
- [64] G. HOEFEL AND C. ELKAN, *Learning a Two-stage SVM/CRF Sequence Classifier*, in *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM '08*, New York, NY, USA, 2008, ACM, pp. 271–278.
- [65] D. K. IAKOVIDIS, D. E. MAROULIS, AND S. A. KARKANIS, *An intelligent system for automatic detection of gastrointestinal adenomas in video endoscopy*, *Computers in Biology and Medicine*, 36 (2006), pp. 1084–1103.
- [66] A. IGNJATOVIC, J. E. EAST, T. GUENTHER, J. HOARE, J. MORRIS, K. RAGUNATH, A. SHONDE, J. SIMMONS, N. SUZUKI, S. THOMAS-GIBSON, AND B. P. SAUNDERS, *What is the most reliable imaging modality for small colonic polyp characterization? study of white-light, autofluorescence, and narrow-band imaging*, *Endoscopy*, 43 (2011), pp. 94–9.
- [67] Y. IMAI, S. KUDO, S. TSURUTA, T. FUJII, S. HAYASHI, AND S. TANAKA, *Problems and clinical significance of v type pit pattern diagnosis: report on round-table consensus meeting*, *Early Colorectal Cancer*, 5 (2001), pp. 595–613.
- [68] C. D. JOHNSON, R. L. MACCARTY, T. J. WELCH, L. A. WILSON, W. S. HARMSSEN, D. M. ILSTRUP, AND D. A. AHLQUIST, *Comparison of the relative sensitivity of CT colonography and double-contrast barium enema for screen detection of colorectal polyps.*, *Clin Gastroenterol Hepatol*, 2 (2004), pp. 314–321.
- [69] R. JONES, *Connected filtering and segmentation using component trees*, *Computer Vision and Image Understanding*, 75 (1999), pp. 215 – 228.
- [70] F. JURIE AND B. TRIGGS, *Creating efficient codebooks for visual recognition*, in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1, Oct 2005, pp. 604–610 Vol. 1.
- [71] H. KANAO, S. TANAKA, S. OKA, M. HIRATA, S. YOSHIDA, AND K. CHAYAMA, *Narrow-band imaging magnification predicts the histology and invasion depth of colorectal tumors.*, *Gastrointest Endosc*, 69 (2009), pp. 631–636.
- [72] S. KARKANIS, D. IAKOVIDIS, D. MAROULIS, D. KARRAS, AND M. TZIVRAS, *Computer-aided tumor detection in endoscopic video using color wavelet features*, *Information Technology in Biomedicine, IEEE Transactions on*, 7 (2003), pp. 141–152.

- [73] J. KARL, N. WILD, M. TACKE, H. ANDRES, U. GARCZAREK, W. ROLLINGER, AND W. ZOLG, *Improved diagnosis of colorectal cancer using a combination of fecal occult blood and novel fecal protein markers.*, Clin Gastroenterol Hepatol, 6 (2008), pp. 1122–1128.
- [74] Z. KATO AND T.-C. PONG, *A markov random field image segmentation model for color textured images*, Image and Vision Computing, 24 (2006), pp. 1103 – 1114.
- [75] R. KIESSLICH, M. GOETZ, M. VIETH, P. R. GALLE, AND M. F. NEURATH, *Technology insight: confocal laser endoscopy for in vivo diagnosis of colorectal cancer.*, Nat Clin Pract Oncol, 4 (2007), pp. 480–490.
- [76] G. KITAGAWA, *Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models*, Journal of Computational and Graphical Statistics, 5 (1996), pp. 1–25.
- [77] T. KOIDE, A. T. HOANG, T. OKAMOTO, S. SHIGEMI, T. MISHIMA, T. TAMAKI, B. RAYTCHEV, K. KANEDA, Y. KOMINAMI, R. MIYAKI, T. MATSUO, S. YOSHIDA, AND S. TANAKA, *Fpga implementation of type identifier for colorectal endoscopic images with nbi magnification*, in Circuits and Systems (APCCAS), 2014 IEEE Asia Pacific Conference on, Nov 2014, pp. 651–654.
- [78] Y. KOMINAMI, S. YOSHIDA, S. TANAKA, Y. SANOMURA, T. HIRAKAWA, B. RAYTCHEV, T. TAMAKI, T. KOIDE, K. KANEDA, AND K. CHAYAMA, *Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy*, Gastrointestinal Endoscopy, 83 (2016), pp. 643–649.
- [79] T. KOSAKA, *Fundamental study on the diminutive polyps of the colon by mucosal stain and dissecting microscope*, Journal of Coloproctology, 28 (1975), pp. 218–228.
- [80] P. KRÄHENBÜHL AND V. KOLTUN, *Efficient inference in fully connected crfs with gaussian edge potentials*, in Advances in Neural Information Processing Systems 24, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, eds., Curran Associates, Inc., 2011, pp. 109–117.
- [81] S. KUDO, S. HIROTA, T. NAKAJIMA, S. HOSOBÉ, H. KUSAKA, T. KOBAYASHI, M. HIMORI, AND A. YAGYUU, *Colorectal tumours and pit pattern.*, Journal of Clinical Pathology, 47 (1994), pp. 880–885.
- [82] S. KUDO, S. TAMURA, T. NAKAJIMA, H.-O. YAMANO, H. KUSAKA, AND H. WATANABE, *Diagnosis of colorectal tumorous lesions by magnifying endoscopy*, Gastrointestinal Endoscopy, 44 (1996), pp. 8–14.

-
- [83] R. KWITT AND A. UHL, *Modeling the Marginal Distributions of Complex Wavelet Coefficient Magnitudes for the Classification of Zoom-Endoscopy Images*, in Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, Oct 2007, pp. 1–8.
- [84] R. KWITT AND A. UHL, *Multi-directional Multi-resolution Transforms for Zoom-Endoscopy Image Classification*, vol. 45, Springer Berlin Heidelberg, 2007, pp. 35–43.
- [85] R. KWITT, A. UHL, M. HÄFNER, A. GANGL, F. WRBA, AND A. VÉCSEI, *Predicting the histology of colorectal lesions in a probabilistic framework*, in Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, June 2010, pp. 103–110.
- [86] U. LADABAUM, A. FIORITTO, A. MITANI, M. DESAI, J. P. KIM, D. K. REX, T. IMPERIALE, AND N. GUNARATNAM, *Real-Time Optical Biopsy of Colon Polyps With Narrow Band Imaging in Community Practice Does Not Yet Meet Key Thresholds for Clinical Decisions*, *Gastroenterology*, 144 (2013), pp. 81–91.
- [87] S. LAZEBNIK, C. SCHMID, AND J. PONCE, *A sparse texture representation using local affine regions*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (2005), pp. 1265–1278.
- [88] S. LAZEBNIK, C. SCHMID, AND J. PONCE, *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories*, in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06), vol. 2, 2006, pp. 2169–2178.
- [89] T. LEUNG AND J. MALIK, *Representing and recognizing the visual appearance of materials using three-dimensional textons*, *International Journal of Computer Vision*, 43 (2001), pp. 29–44.
- [90] W. LI, U. GUSTAFSSON, AND Y. A-RAHIM, *Automatic colonic lesion detection and tracking in endoscopic videos*, in Proc. SPIE, vol. 7963, 2011, pp. 79632L–79632L–8.
- [91] M. LIN, K. WONG, W. L. NG, I. H. SHON, AND M. MORGAN, *Positron emission tomography and colorectal cancer.*, *Crit Rev Oncol Hematol*, 77 (2011), pp. 30–47.
- [92] C. LIU, J. YUEN, AND A. TORRALBA, *Nonparametric scene parsing: Label transfer via dense scene alignment*, *Artificial Intelligence*, (2009).
- [93] F. LIU, G. LIN, AND C. SHEN, *CRF learning with CNN features for image segmentation*, *Pattern Recognition*, 48 (2015), pp. 2983 – 2992.
- [94] G. LIU, G. S. XIA, W. YANG, AND L. ZHANG, *Texture analysis with shape co-occurrence patterns*, in Pattern Recognition (ICPR), 2014 22nd International Conference on, Aug 2014, pp. 1627–1632.

- [95] J. LIU, K. R. SUBRAMANIAN, AND T. S. YOO, *A robust method to track colonoscopy videos with non-informative images*, Int J Comput Assist Radiol Surg, 8 (2013), pp. 575–92.
- [96] R. LIU, Z. LI, AND J. JIA, *Image partial blur detection and classification*, in Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, June 2008, pp. 1–8.
- [97] J. LONG, E. SHELHAMER, AND T. DARRELL, *Fully convolutional networks for semantic segmentation*, in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 3431–3440.
- [98] D. G. LOWE, *Object recognition from local scale-invariant features*, in Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 2, 1999, pp. 1150–1157 vol.2.
- [99] D. G. LOWE, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision, 60 (2004), pp. 91–110.
- [100] H. MACHIDA, Y. SANO, Y. HAMAMOTO, M. MUTO, T. KOZU, H. TAJIRI, AND S. YOSHIDA, *Narrow-band imaging in the diagnosis of colorectal mucosal lesions: a pilot study.*, Endoscopy, 36 (2004), pp. 1094–1098.
- [101] S. G. MALLAT, *Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbb{R})$* , Transactions of the American mathematical society, 315 (1989), pp. 69–87.
- [102] S. G. MALLAT, *A theory for multiresolution signal decomposition: the wavelet representation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 11 (1989), pp. 674–693.
- [103] S. MANIVANNAN, R. WANG, E. TRUCCO, AND A. HOOD, *Automatic normal-abnormal video frame classification for colonoscopy*, in Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on, April 2013, pp. 644–647.
- [104] S. MANIVANNAN, R. WANG, M. TRUJILLO, J. HOYOS, AND E. TRUCCO, *Video-Specific SVMs for Colonoscopy Image Classification*, in Computer-Assisted and Robotic Endoscopy, X. Luo, T. Reichl, D. Mirota, and T. Soper, eds., Lecture Notes in Computer Science, Springer International Publishing, 2014, pp. 11–21.
- [105] D. E. MAROULIS, D. K. IAKOVIDIS, S. A. KARKANIS, AND D. A. KARRAS, *CoLD: a versatile detection system for colorectal lesions in endoscopy video-frames*, Comput Methods Programs Biomed, 70 (2003), pp. 151–66.
- [106] B. MAYINGER, Y. OEZTURK, M. STOLTE, G. FALLER, J. BENNINGER, D. SCHWAB, J. MAISS, E. G. HAHN, AND S. MUEHLDOERFER, *Evaluation of sensitivity and inter-*

- and intra-observer variability in the detection of intestinal metaplasia and dysplasia in Barrett's esophagus with enhanced magnification endoscopy*, Scand J Gastroenterol, 41 (2006), pp. 349–56.
- [107] A. MEINING, T. RÖSCH, R. KIESSLICH, M. MUDERS, F. SAX, AND W. HELDWEIN, *Inter- and intra-observer variability of magnification chromoendoscopy for detecting specialized intestinal metaplasia at the gastroesophageal junction*, Endoscopy, 36 (2004), pp. 160–4.
- [108] MINISTRY OF HEALTH, LABOR AND WELFARE, *The general situation of demographic statistics in 2015*, <http://www.mhlw.go.jp/toukei/saikin/hw/jinkou/kakutei15/index.html>, 2015, [Online; accessed 2017/03/02].
- [109] P. MONASSE AND F. GUICHARD, *Scale-space from a level lines tree*, Journal of Visual Communication and Image Representation, 11 (2000), pp. 224 – 236.
- [110] G. MOSER AND S. SERPICO, *Contextual remote-sensing image classification by support vector machines and Markov random fields*, in Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International, July 2010, pp. 3728–3731.
- [111] G. MOSER AND S. B. SERPICO, *Combining support vector machines and markov random fields in an integrated framework for contextual image classification*, IEEE Transactions on Geoscience and Remote Sensing, 51 (2013), pp. 2734–2752.
- [112] L. NAJMAN AND M. COUPRIE, *Building the component tree in quasi-linear time*, IEEE Transactions on Image Processing, 15 (2006), pp. 3531–3539.
- [113] A. NARAYANAN, *Algorithm as 266: Maximum likelihood estimation of the parameters of the dirichlet distribution*, Journal of the Royal Statistical Society. Series C (Applied Statistics), 40 (1991), pp. pp. 365–374.
- [114] NATIONAL CANCER INSTITUTE, *Cancer stat facts: Colon and rectum cancer*, <https://seer.cancer.gov/statfacts/html/colorect.html>, 2016, [Online; accessed 2017/03/02].
- [115] M.-E. NILSBACK AND A. ZISSERMAN, *Delving deeper into the whorl of flower segmentation*, Image and Vision Computing, 28 (2010), pp. 1049 – 1062.
- [116] D. NISTER AND H. STEWENIUS, *Scalable recognition with a vocabulary tree*, in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, 2006, pp. 2161–2168.
- [117] H. NOH, S. HONG, AND B. HAN, *Learning deconvolution network for semantic segmentation*, in 2015 IEEE International Conference on Computer Vision (ICCV), Dec 2015, pp. 1520–1528.

- [118] E. NOWAK, F. JURIE, AND B. TRIGGS, *Sampling Strategies for Bag-of-Features Image Classification*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 490–503.
- [119] S. OBA, S. TANAKA, S. OKA, H. KANAO, S. YOSHIDA, F. SHIMAMOTO, AND K. CHAYAMA, *Characterization of colorectal tumors using narrow-band imaging magnification: combined diagnosis with both pit pattern and microvessel features.*, Scand J Gastroenterol, 45 (2010), pp. 1084–1092.
- [120] J. OH, S. HWANG, J. LEE, W. TAVANAPONG, J. WONG, AND P. C. DE GROEN, *Informative frame classification for endoscopy video*, Medical Image Analysis, 11 (2007), pp. 110–127.
- [121] T. OJALA, M. PIETIKAINEN, AND T. MAENPAA, *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24 (2002), pp. 971–987.
- [122] A. OTO, *Virtual endoscopy.*, Eur J Radiol, 42 (2002), pp. 231–239.
- [123] A. R. PADHANI, *Advances in imaging of colorectal cancer.*, Crit Rev Oncol Hematol, 30 (1999), pp. 189–199.
- [124] S. Y. PARK, D. SARGENT, I. SPOFFORD, K. G. VOSBURGH, AND Y. A-RAHIM, *A colon video analysis framework for polyp detection*, IEEE transactions on bio-medical engineering, 59 (2012), pp. 1408–1418.
- [125] S. G. PATEL, P. SCHOENFELD, H. M. KIM, E. K. WARD, A. BANSAL, Y. KIM, L. HOSFORD, A. MYERS, S. FOSTER, J. CRAFT, S. SHOPINSKI, R. H. WILSON, D. J. AHNEN, A. RASTOGI, AND S. WANI, *Real-Time Characterization of Diminutive Colorectal Polyp Histology Using Narrow-Band Imaging: Implications for the Resect and Discard Strategy*, Gastroenterology, 150 (2016), pp. 406–418.
- [126] B. PERRET AND C. COLLET, *Connected image processing with multivariate attributes: An unsupervised markovian classification approach*, Computer Vision and Image Understanding, 133 (2015), pp. 1 – 14.
- [127] F. PROVOST AND R. KOHAVI, *On Applied Research in Machine Learning*, in Machine learning, 1998, pp. 127–132.
- [128] A. RASTOGI, D. S. RAO, N. GUPTA, S. W. GRISOLANO, D. C. BUCKLES, E. SIDORENKO, J. BONINO, T. MATSUDA, E. DEKKER, T. KALTENBACH, R. SINGH, S. WANI, P. SHARMA, M. S. OLYAEE, A. BANSAL, AND J. E. EAST, *Impact of a computer-based teaching module on characterization of diminutive colon polyps by using narrow-band imaging by non-experts in academic and community practice: a video-based study*, Gastrointestinal Endoscopy, 79 (2014), pp. 390–398.

- [129] D. K. REX, C. KAHN, M. O'BRIEN, T. LEVIN, H. POHL, A. RASTOGI, L. BURGART, T. IMPERIALE, U. LADABAUM, J. COHEN, AND D. A. LIEBERMAN, *The American Society for Gastrointestinal Endoscopy PIVI (Preservation and Incorporation of Valuable Endoscopic Innovations) on real-time endoscopic assessment of the histology of diminutive colorectal polyps*, *Gastrointestinal Endoscopy*, 73 (2011), pp. 419–422.
- [130] B. RISTIC, S. ARULAMPALAM, AND N. GORDON, *Beyond the Kalman filter : particle filters for tracking applications*, Artech House, Boston, London, 2004.
- [131] P. SALEMBIER AND L. GARRIDO, *Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval*, *IEEE Transactions on Image Processing*, 9 (2000), pp. 561–576.
- [132] K. W. SANFORD AND R. A. MCPHERSON, *Fecal Occult Blood Testing*, *Clinics in laboratory medicine*, 29 (2009), pp. 523–541.
- [133] Y. SANO, T. HORIMATSU, K. I. FU, A. KATAGIRI, M. MUTO, AND H. ISHIKAWA, *Magnifying observation of microvascular architecture of colorectal lesions using a narrow-band imaging system*, *Digestive Endoscopy*, 18 (2006), pp. S44–S51.
- [134] B. SCHÖLKOPF AND A. J. SMOLA, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT press, 2002.
- [135] J. SERRA, *Image Analysis and Mathematical Morphology*, Academic Press, Inc., Orlando, FL, USA, 1983.
- [136] L. K. SHIN, P. POULLOS, AND R. B. JEFFREY, *MR colonography and MR enterography.*, *Gastrointest Endosc Clin N Am*, 20 (2010), pp. 323–346.
- [137] J. SHOTTON, J. WINN, C. ROTHER, AND A. CRIMINISI, *TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 1–15.
- [138] J. SHOTTON, J. WINN, C. ROTHER, AND A. CRIMINISI, *Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context*, *International Journal of Computer Vision*, 81 (2009), pp. 2–23.
- [139] S. SONOYAMA, T. HIRAKAWA, T. TAMAKI, T. KURITA, B. RAYTCHEV, K. KANEDA, T. KOIDE, S. YOSHIDA, Y. KOMINAMI, AND S. TANAKA, *Transfer learning for bag-of-visual words approach to nbi endoscopic image classification*, in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2015, pp. 785–788.

- [140] S. SONOYAMA, T. TAMAKI, T. HIRAKAWA, B. RAYTCHEV, K. KANEDA, T. KOIDE, S. YOSHIDA, H. MIENO, AND S. TANAKA, *Transfer learning for endoscopic image classification*, in Proceedings of The Korea-Japan joint workshop on Frontiers of Computer Vision, FCV2016, 2016.
- [141] T. STEHLE, R. AUER, S. GROSS, A. BEHRENS, J. WULFF, T. AACH, R. WINOGRAD, C. TRAUTWEIN, AND J. TISCHENDORF, *Classification of colon polyps in NBI endoscopy using vascularization features*, in Proc. SPIE, vol. 7260, 2009, pp. 72602S–72602S–12.
- [142] I. STEINWART AND A. CHRISTMANN, *Support Vector Machines*, New York, Springer, first ed., 2008.
- [143] Y. TAKEMURA, S. YOSHIDA, S. TANAKA, R. KAWASE, K. ONJI, S. OKA, T. TAMAKI, B. RAYTCHEV, K. KANEDA, M. YOSHIHARA, AND K. CHAYAMA, *Computer-aided system for predicting the histology of colorectal tumors by using narrow-band imaging magnifying colonoscopy (with video)*, *Gastrointest Endosc*, 75 (2012), pp. 179–85.
- [144] Y. TAKEMURA, S. YOSHIDA, S. TANAKA, K. ONJI, S. OKA, T. TAMAKI, K. KANEDA, M. YOSHIHARA, AND K. CHAYAMA, *Quantitative analysis and development of a computer-aided system for identification of regular pit patterns of colorectal lesions*, *Gastrointest Endosc*, 72 (2010), pp. 1047–51.
- [145] T. TAMAKI, 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 (2013), pp. 78 – 100.
- [146] T. TAMAKI, J. YOSHIMUTA, T. TAKEDA, B. RAYTCHEV, K. KANEDA, S. YOSHIDA, Y. TAKEMURA, AND S. TANAKA, *A system for colorectal tumor classification in magnifying endoscopic nbi images*, in Asian Conference on Computer Vision (ACCV), vol. 6493, Springer, 2010, pp. 452–463.
- [147] S. TANAKA, T. KALTENBACH, K. CHAYAMA, AND R. SOETIKNO, *High-magnification colonoscopy (with videos)*, *Gastrointestinal endoscopy*, 64 (2006), pp. 604–613.
- [148] W. TANG, Y. WANG, AND W. HE, *An image segmentation algorithm based on improved multiscale random field model in wavelet domain*, *Journal of Ambient Intelligence and Humanized Computing*, 7 (2016), pp. 221–228.
- [149] J. TIGHE AND S. LAZEBNIK, *SuperParsing: Scalable Nonparametric Image Parsing with Superpixels*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 352–365.
- [150] J. TIGHE AND S. LAZEBNIK, *Superparsing*, *International Journal of Computer Vision*, 101 (2013), pp. 329–349.

-
- [151] J. J. W. TISCHENDORF, S. GROSS, R. WINOGRAD, H. HECKER, R. AUER, A. BEHRENS, C. TRAUTWEIN, T. AACH, AND T. STEHLE, *Computer-aided classification of colorectal polyps based on vascular patterns: a pilot study*, *Endoscopy*, 42 (2010), pp. 203–7.
- [152] M. R. TURNER, *Texture discrimination by gabor functions*, *Biological Cybernetics*, 55 (1986), pp. 71–82.
- [153] V. N. VAPNIK AND V. VAPNIK, *Statistical learning theory*, vol. 1, Wiley New York, 1998.
- [154] A. VEDALDI AND B. FULKERSON, *Vlfeat: An open and portable library of computer vision algorithms*, in *Proceedings of the 18th ACM international conference on Multimedia*, ACM, 2010, pp. 1469–1472.
- [155] A. VOISIN, V. A. KRYLOV, G. MOSER, S. B. SERPICO, AND J. ZERUBIA, *Classification of very high resolution sar images of urban areas using copulas and texture in a hierarchical markov random field model*, *IEEE Geoscience and Remote Sensing Letters*, 10 (2013), pp. 96–100.
- [156] L. WANG AND B. MANJUNATH, *A semantic representation for image retrieval*, in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 2, Sept 2003, pp. II–523–6 vol.3.
- [157] T. WU, M. H. BAE, M. ZHANG, R. PAN, AND A. BADEA, *A prior feature SVM-MRF based method for mouse brain segmentation*, *NeuroImage*, 59 (2012), pp. 2298–2306.
- [158] G.-S. XIA, J. DELON, AND Y. GOUSSEAU, *Shape-based invariant texture indexing*, *International Journal of Computer Vision*, 88 (2010), pp. 382–403.
- [159] Y. XU, T. GÉRAUD, AND L. NAJMAN, *Two Applications of Shape-Based Morphology: Blood Vessels Segmentation and a Generalization of Constrained Connectivity*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 390–401.
- [160] H. YOSHIDA AND A. H. DACHMAN, *Computer-aided diagnosis for CT colonography.*, *Semin Ultrasound CT MR*, 25 (2004), pp. 419–431.
- [161] J. ZHANG, M. MARSZALEK, S. LAZEBNIK, AND C. SCHMID, *Local features and kernels for classification of texture and object categories: A comprehensive study*, *International Journal of Computer Vision*, 73 (2007), pp. 213–238.
- [162] Y. ZHANG, M. BRADY, AND S. SMITH, *Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm*, *IEEE Transactions on Medical Imaging*, 20 (2001), pp. 45–57.

