Robust classification and detection with applications in biomedical images

Lin, Dongyun

2018

Lin, D. (2018). Robust classification and detection with applications in biomedical images.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/90200>

<https://doi.org/10.32657/10220/47324>

Robust Classification and Detection with Applications in Biomedical Images

LIN DONGYUN

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2018

Statement of originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

SIGNED: *Lin Dongyun* DATE: 26/12/2018

Acknowledgment

Foremost, I would like to express my deepest gratitude to my supervisor, Associate Professor Lin Zhiping and co-supervisor, Professor Liu Ai Qun for their constant support, professional guidance and continuous encouragement through my four years of Ph.D. study and research. Without their insightful comments and constructive suggestions, I would not have completed my Ph.D. research and this thesis. Their insightful, passionate and strict attitude towards research works will continuously benefit me throughout my lifetime.

I would also show my gratitude to Professor Kar-Ann Toh, Associate Professor Sun Lei and Dr. Zhang Jing Bo for providing great help and support to my research works. I sincerely thank the technicians from the Center for Bio Devices and Signal Analysis for their technical support during my Ph.D. study. I would like to thank Nanyang Technological University for the four years financial support of my Ph.D. scholarship.

During the years of my Ph.D. study, I have been in the company of numerous wonderful friends. I have enjoyed spending my time with Mr. Liang Yunfeng, Mr. Peng Siyuan, and Mr. Zhuang Huiping. I benefit a lot from their help in my daily life.

Finally, and most importantly, I would like to thank my beloved family for the endless love and unconditional support throughout the journey of my study.

Abstract

In this thesis, robust classification and detection methods with applications in biomedical images have been studied. The study is motivated by Chow's theoretical works on classification with a reject option and specific applications in biomedical images. Technically, the study has been carried out from two perspectives. The first is proposing robust classification methods with a reject option based on Chow's theoretical works to select ambiguous and outlier samples. The second is proposing robust multi-stage methods for biomedical image classification and detection.

Firstly, we propose a twin support vector machine with a reject option (RO-TWSVM) through the Receiver Operating Characteristic (ROC) curve for binary classification. The method is formulated under a cost-sensitive framework and follows the principle of minimization of the expected classification cost. Extensive experiments are conducted on synthetic and real-world datasets to compare the proposed RO-TWSVM with the original TWSVM without a reject option (TWSVM-without-RO) and the existing SVM with a reject option (RO-SVM). The experimental results demonstrate that our RO-TWSVM significantly outperforms TWSVM-without-RO, and in general, performs better than RO-SVM.

Secondly, we propose a novel two-stage method for robust biomedical image classification. This method is proposed based on a cascade of a support vector machine (SVM) with a reject option and subspace analysis to robustly classify biomedical images. At stage 1, we propose a confidence score based classification rule with a reject option for a preliminary decision using the SVM. The testing images going through stage 1

are separated into two groups based on their confidence scores. Those testing images with sufficiently high confidence scores are classified at stage 1 while the others with low confidence scores are rejected and fed to stage 2. At stage 2, the rejected images from stage 1 are first processed by a subspace analysis technique called eigenfeature regularization and extraction (ERE), and then classified by another SVM trained in the transformed subspace learned by ERE. Extensive experiments implemented on three benchmark datasets show that the proposed method significantly outperforms several state-of-the-art competing methods in terms of multiple classification performance metrics.

Thirdly, we propose a novel two-stage method for robust biomedical image detection. We investigate a specific task of detecting ring-like endosomes in fluorescent microscopy images. The proposed method first selects candidate patches using local feature information. Then, it identifies endosomes from candidate patches using a pretrained SVM based on global information of training patches. The experiments conducted on real-world microscopy images show that the proposed method significantly outperforms several state-of-the-art competing methods in terms of multiple detection performance metrics.

Overall, the thesis contributes two types of novel methods to enhance robustness of classification and detection methods: (i) Proposing a twin support vector machine with a reject option (RO-TWSVM) through the Receiver Operating Characteristic (ROC) curve to select ambiguous and outlier testing samples; (ii) Proposing multi-stage methods for the classification or detection of biomedical images.

Contents

Statement of originality	i
Acknowledgement	i
Abstract	iii
List of Figures	viii
List of Tables	xiv
List of Acronyms	xvi
1 Introduction	1
1.1 Background and motivation	2
1.2 Objectives and contributions	4
1.3 Thesis organization	7
2 Literature review	9
2.1 Classification with a reject option	10
2.1.1 Chow's optimal classification rule with a reject option	10
2.1.2 Support vector machine with a reject option using the ROC curve (RO-SVM)	13
2.2 Twin support vector machine (TWSVM)	19
2.3 Biomedical image feature extraction	22

2.3.1	Local feature based bag-of-words (BoW) model	22
2.3.2	Eigenfeature regularization and extraction	28
2.4	Applications in biomedical images	31
2.4.1	Biomedical image classification	31
2.4.2	Biomedical image detection	32
3	Twin SVM with a reject option through ROC curve (RO-TWSVM)	37
3.1	Introduction	38
3.2	Methodology	39
3.3	Experiments on synthetic datasets	42
3.3.1	Synthetic data 1: uniformly distributed data with Gaussian noise	43
3.3.2	Synthetic data 2: “Cross Planes” data	47
3.4	Experiments on real-world benchmark datasets	51
3.4.1	Wilcoxon rank sum test	51
3.4.2	Friedman and Nemenyi post-hoc test	56
3.5	Summary	63
4	A two-stage method for robust biomedical image classification	65
4.1	Introduction	66
4.2	Methodology	67
4.2.1	Overview of the proposed method	67
4.2.2	Stage 1: Support vector machine with a reject option	68
4.2.3	Stage 2: Subspace analysis of the rejected images	74
4.3	Experiments	76
4.3.1	Datasets	76
4.3.2	Experimental Settings	79
4.3.3	Experiment 1	80
4.3.4	Experiment 2	87
4.4	Summary	91

5	A two-stage method for robust ring-like endosome detection	93
5.1	Introduction	94
5.2	Methodology	97
5.2.1	Feature extraction	97
5.2.2	Localization stage	97
5.2.3	Identification stage	103
5.3	Experiments	105
5.3.1	Dataset description	108
5.3.2	Performance metrics	108
5.3.3	Experimental protocol and parameter settings	110
5.3.4	Experimental results and analysis	111
5.4	Summary	116
6	Conclusions and future works	118
6.1	Conclusions	119
6.2	Future works	122
	Bibliography	126
	Author's publications	

List of Figures

2.1	The case when condition (2.24) is satisfied and two level curves do not intersect at the same point with the convex hull of the ROC curve. . . .	17
2.2	The case when condition (2.24) is violated.	18
2.3	The case when condition (2.24) is satisfied but two level curves intersect at the same point with the convex hull of the ROC curve.	18
2.4	An illustration to show the differences between the SVM hyperplane and the TWSVM hyperplanes. (a) The SVM hyperplane; (b) The TWSVM hyperplanes.	21
2.5	An example of extracting dense SIFT features from a microscopy image.	23
2.6	(a) The training images are decomposed into many SIFT features. By K-means clustering, these SIFT features are clustered to a predefined number of classes. The centroids of these clusters are referred to as the visual words. In this example, three visual words are generated (shown by black crosses). The set containing all the visual words is referred to as a dictionary. (b) Given an image to be represented by a bag-of-words model, it is first decomposed into several SIFT features. Each SIFT feature is assigned to its nearest visual word using Euclidean distance. Implementing such feature assignment for each SIFT feature, the image can be represented as a histogram of visual word frequencies, i.e., the counts of the occurrences of the SIFT features assigned to each visual word.	27

3.1	The reject region for a 2D example. The blue dashed lines delimit the rejection boundaries and the black dashed-dotted lines represent nonparallel hyperplanes trained by TWSVM.	40
3.2	(a) Synthetic dataset 1, (b) The reject region boundaries of RO-TWSVM for each cost model: C3 (black solid line), C4(blue dashed line), C5 (cyan dotted line) and C6 (magenta dashed-dotted line).	44
3.3	Approximated class-conditional probability density functions of (a) Distance difference $d(\mathbf{x})$ of TWSVM and (b) SVM outputs for synthetic dataset 1.	46
3.4	(a) “Cross Planes” dataset. (b) The reject region generated by the proposed RO-TWSVM. (c) The reject region generated by RO-SVM.	48
3.5	Approximated class-conditional probability density functions of (a) Distance difference $d(\mathbf{x})$ of TWSVM and (b) SVM outputs for “Cross Planes” dataset.	50
3.6	The average cost for each dataset : (a) linear kernel , (b) RBF kernel.	59
3.7	Comparison results of four methods using linear kernel through Nemenyi test: (a) CM1, (b) CM2, (c) CM3 and (d) CM4.	60
3.8	Comparison results of 4 methods using RBF kernel through Nemenyi test: (a) CM1, (b) CM2, (c) CM3 and (d) CM4.	61
4.1	The classification flowchart of the proposed method for a batch of testing images.	68

- 4.2 A diagram illustrating an example which contains 3 categories of samples shown by the blue triangles, red circles and purple squares, respectively. In this figure, we only plot the hyperplane to discriminate the samples of the category shown by the blue triangles. The signed distance D_h is defined as the perpendicular distance between a testing sample (denoted by the green diamond) and the SVM hyperplane with the sign determined by the SVM output. D_h is adopted in the calculation of the discriminative score for blue triangle category. The discriminative scores for the other category can be calculated in a similar way. $D_c^m(m = 1, 2, 3)$ denotes the distance between the testing sample and the m_{th} category centroid and they are applied to calculate the prior scores. . . . 70
- 4.3 The figure shows the error rate with rejection (blue dashed line), the rejection rate (red solid line) and the error rate without reject option (grey dotted line) with respect to t for the Hep-2 cell image dataset. 73
- 4.4 The procedure of stage 1. At this stage, the SIFT feature based BoW model is applied. In the training phase, an SVM with a reject option is trained. In the testing phase, the testing images are either classified with high confidence or rejected and fed to stage 2. 73
- 4.5 The 2D scatter plots for the Hep-2 cell image dataset with 6 categories of images: (a) the samples in the PCA transformed subspace. (b) the samples in the ERE transformed subspace. 75
- 4.6 The procedure of stage 2. At this stage, the SURF feature based BoW model is applied. In the training phase, ERE is applied on the training samples and an SVM is trained in the transformed subspace. In the testing phase, the rejected samples (images) from stage 1 are transformed to the ERE learned subspace and classified by the SVM. . . 76
- 4.7 Some sample images in the 2D Hela dataset for each category. 77
- 4.8 Some sample images in the PAP smear dataset for each category. 78

4.9	Some sample images in the Hep-2 cell dataset for each category.	79
4.10	The classification accuracies from 30 rounds of experiments for each dataset : (a) 2D Hela, (b) PAP and (3) Hep-2. The blue dashed lines denote the accuracies of the method of stage 1 and the red dotted lines show that of the method of stage 2. The green solid lines denote the accuracies of the proposed method.	82
4.11	The ROC curves for the proposed method (the blue solid line), the method of stage 1 (the red dotted line) and the method of stage 2 (the black dashed-dotted line), respectively.	83
4.12	An illustration of the classification of two biomedical patterns from 2D Hela dataset : (a) an image with high geometric correspondence. (b) an image without high geometric correspondence.	86
4.13	The confusion matrices for the 2D Hela dataset generated by (a) the proposed method; (b) the best SAHLBP based method. C1: Actin; C2: DNA; C3: Endosome; C4: Er; C5: Golgia; C6: Golgpp; C7: Lysosome; C8: Mircrotubules; C9: Mitochondiria; C10: Nucleolus.	88
4.14	The confusion matrices for the Hep-2 cell dataset generated by (a) the proposed method; (b) the best SAHLBP based method. C1: Centromere; C2: Coarse speckled; C3: Cytoplasmatic; C4: Fine speckled; C5: Homogeneous; C6: Nucleolar.	90
5.1	(a) A microscopy image of a cell with multiple subcellular structures. (b) Three patches of ring-like endosomes. (c) Three patches of non-endosome background patterns. The scale bar for (a) is 5 μm and that for (b) and (c) is 1 μm	94
5.2	A microscopy image with manual annotations (using yellow rectangles) of ring-like endosomes in a microscopy image.	95

- 5.3 (a) The candidate patch of an endosome in a microscopy image can be determined by the centroid of the endosome (denoted by the red dot) and a rectangle box which contains the whole endosome (denoted by dashed green rectangles). (b) The voting-map for the microscopy image shown in (a). Scale bar = 1 μm 98
- 5.4 A diagram illustrating the process of voting-map generation. Three query patches and two scaling factors , i.e., 0.8 and 1, are adopted in this example. (a) The testing microscopy image. (b) The three query patches. (c) The generation of the voting-maps with the scaling factor set at 0.8. For each query patch, a voting-map is generated by local feature matching between the query patch and the testing image. Then a voting-map associated with the scaling factor of 0.8 is obtained by summing the three voting-maps. (d) The generation of the voting-maps with the scaling factor set at 1. (e) The final voting-map is determined by selecting the optimal scaling factor for each pixel position as described in the text. Scale bar = 1 μm 101
- 5.5 A diagram illustrating two cases of local feature matching in the voting-map generation procedure. 103
- 5.6 A diagram illustrating the identification stage using SVM. (a) After the voting-map thresholding and non-maximum suppression, the candidate patches of endosomes are obtained. (b) Each candidate patch is represented by an LLC-based bag-of-words histogram. (c) The histogram is classified by the SVM to determine whether the patch contains an endosome or a non-endosome background pattern. Scale bar = 1 μm . . . 107

- 5.7 The detection results for an example microscopy image with the total magnification 100.8X using: (a) the proposed method, (b) AB-Haar, (c) Pixel-LDA, (d) MO-FRST. The true positives, the false positives and the false negatives are represented by gree, red, and yellow patches, respectively in the color version of this thesis. In (a), we show three enlarged patches which are a false positive (indicated by a red arrow), a true positive (green arrow) and a false negative (yellow arrow). The scale bars for the big microscopy images is $5\text{ }\mu\text{m}$ and that for the three small patches shown in (a) is $1\text{ }\mu\text{m}$ 112
- 5.8 The detection results for another example microscopy image with the total magnification 157.5X using: (a) the proposed method, (b) AB-Haar, (c) Pixel-LDA, (d) MO-FRST. Here, the notations are the same as Fig. 5.7. Scale bar = $5\text{ }\mu\text{m}$ 113
- 5.9 The MR-FPPI curves of the compared methods for (a) the microscopy images with the total magnification of 100.8X, (b) the microscopy images with the total magnification of 157.5X. 115

List of Tables

2.1	Cost matrix.	15
3.1	The cost models for synthetic data 1.	43
3.2	The expected costs for different cost models for synthetic data 1.	45
3.3	The expected costs for different cost models for “Cross Planes” dataset.	47
3.4	Summary of experimental datasets for Wilcoxon rank sum test.	51
3.5	The cost models for the experiments on real-world datasets.	51
3.6	Comparison between RO-TWSVM and TWSVM-without-RO for linear kernel based on Wilcoxon rank sum test.	54
3.7	Comparison between RO-TWSVM and TWSVM-without-RO for RBF kernel based on Wilcoxon rank sum test.	54
3.8	Comparison between RO-TWSVM and RO-SVM for linear kernel based on Wilcoxon rank sum test.	55
3.9	Comparison between RO-TWSVM and RO-SVM for RBF kernel based on Wilcoxon rank sum test.	56
3.10	Summary of experimental datasets for Friedman and Nemenyi tests.	58
3.11	Summary of the experimental results	62
4.1	A summary of the 2D Hela dataset	77
4.2	A summary of the PAP smear dataset	78
4.3	A summary of the Hep-2 cell image dataset	79

4.4	Classification accuracy comparison of the proposed two-stage method and the individual method of each stage.	81
4.5	The AUC, TPR and FPR for the PAP smear dataset	83
4.6	The comparison between our two-stage method and the method with a direct fusion of the two features	84
4.7	Classification accuracy comparison among the proposed method and the SAHLBP based methods	87
4.8	Classification accuracy comparison among the proposed method and the competing methods under the bag-of-words framework	91
5.1	The dataset information	108
5.2	Average performance metrics (and standard deviations) for the compared methods	114

List of Acronyms

AB-Haar Adaboost on Haar-like features

BoW Bag-of-Words

CD Critical Difference

CFN Cost of False Negative

CFP Cost of False Positive

CM Cost Model

CR Cost of Rejection

CS Confidence Score

CTN Cost of True Negative

CTP Cost of True Positive

DS Discriminative Score

EC Expected Cost

ERE Eigenfeature Regularization and Extraction

FNR False Negative Rate

FPPI False Positives Per Images

FPR False Positive Rate

FRST Fast Radial Symmetry Transformation

LBP Local Binary Pattern

LDA-Pixel Linear Discriminative Analysis on Pixel based features

LLC Locally-constrained Linear Coding

MO-FRST Morphological Operations Combined with Fast Radial Symmetry Transformation

MR Miss Rate

PCA Principle Component Analysis

PS Prior Score

RO Reject Option

RO-SVM Support Vector Machine with a Reject Option

RO-TWSVM Twin Support Vector Machine with a Reject Option

ROC Receiver Characteristic Curve

SAHLBP Spatial Adjacent Histogram Based on Adaptive Local Binary Pattern

SIFT Scale Invariant Feature Transform

SPM Spatial Pyramid Matching

SURF Speeded-UP Robust Features

SVM Support Vector Machine

TNR True Negative Rate

TPR True Positive Rate

TWSVM Twin Support Vector Machine

Chapter 1

Introduction

In this chapter, the background and motivation of this thesis are first described in Section 1.1. In Section 1.2, the major objectives and contributions are summarized. Finally, the organization of the remaining parts of the thesis is presented in Section 1.3.

1.1 Background and motivation

In real-world applications, it is highly desired to build robust machine learning methods to perform classification or detection with high confidence over uncertainties [1]. Essentially, the goal of a classification task is to predict the identity (or class label) of a testing sample. A detection task, on the other hand, aims at not only predicting the identity of a testing sample but also determining the location of the sample. For example, in human detection [2], a detection method searches in a testing image to first locate relatively small-sized patches, i.e., the rectangular bounding boxes, and then checks whether each of these patches contains human beings.

Traditional supervised classification tasks are addressed by a paradigm consisting of two phases: (i) In the training phase, a classification model is trained using a number of training samples (data) and their categorical information (class labels); (ii) In the testing phase, a new testing sample is fed to the model and its class label is predicted. This paradigm is also applied in a detection task as the classification is usually required. In a real-world classification or detection task, two types of uncertain testing samples, namely ambiguous and outlier samples, are potential to cause errors in classification or detection. Ambiguous samples are those which are similar to several classes and appear at the decision boundary generated by a classifier. Outlier samples are those which have no similarity to any of the training classes. Outlier samples are very common for a detection task. For example, in human detection, those patches containing non-human objects and not being seen in the training phase could be treated as outlier samples. The sources of these two uncertain samples could be: (i) data flaws, e.g., wrong measurement in some data features; (ii) limitations of classification or detection methods, e.g., poor data representations. Therefore, it is valuable to study the robustness of classification or detection methods with respect to ambiguous or outlier testing samples from these two sources. Towards (i), we investigate the ways to select ambiguous or outlier samples. For (ii), we investigate the ways to alleviate the limitations of some classification or detection methods with applications in biomedical images

The motivation of this thesis is twofold. The first is inspired by a classic theoretical work in early machine learning literature [3, 4] where a reject option is embedded into classification methods to select testing samples with low maximum posterior probabilities. The maximum posterior probability is a confidence measurement for the classification decision. Unlike standard classifiers which assign each testing sample to a specific class, classifiers with a reject option withhold or abstain from making the decision if the sample is in a region with low confidence. Such a region is called the reject region. Classifiers with a reject option are thus called abstaining classifiers [5] or selective classifiers [6]. Since both ambiguous and outlier samples are the samples with low maximum posterior probabilities, they could be selected by Chow's rule.

Chow's rule achieves great success in practical applications. Classifiers with a reject option are adopted in many applications including biometric verification [7], automated medical diagnosis [8], genetic engineering [9], text categorization [10] and image categorization [11] etc. In practice, if a sample is labeled by a classifier with very low confidence, it is appropriate to identify it as 'rejected' rather than assigning it with a categorical label. This is particularly useful in automated medical diagnosis where the consequence of misclassification can be very harmful [8]. For instance, a wrong diagnosis of a terminal illness in either way could both lead to a disastrous reaction of a patient. For such cases, additional human investigations or observations can be conducted to avoid misclassification. Motivated by Chow's theory and aforementioned practical applications, we investigate ways to exploit classification with a reject option to effectively select both ambiguous and outlier samples (towards the first source of uncertain testing samples). In this thesis, the first topic is proposing novel classification methods with a reject option to select ambiguous or outlier testing samples.

The second motivation is from specific applications in biomedical images. Biomedical images are playing significant roles in modern biological studies and medical diagnosis processes. Manual classification or detection of biomedical images is prone to subjectivity and suffers from inefficiency due to high-throughput production of images

in practice. Hence, it is greatly demanded for automating the process of biomedical image classification or detection using machine learning approaches. In the last decade, machine learning approaches have achieved great success towards applications in biomedical images (see e.g., [12] for a general review). For instance, the structures and functions of different proteins were studied based on their locations in specific organelles using image classification models [13–16]. In [17, 18], mixture patterns of protein distributions were recognized and [19] proposed to identify protein distribution patterns in time series images. In [20, 21], generative models were trained from microscopy images to capture variations of protein distributions from cell to cell.

Automated biomedical image classification or detection could confront challenges of high level noise, image blur, illumination variation and complicated geometric correspondence among various categorical biomedical patterns, etc [12, 22]. Not surprisingly, ambiguous and outlier samples are frequently encountered in tasks of biomedical image classification or detection. For applications in biomedical images, ambiguous and outlier samples mainly come from the second source, i.e., limitations of classification or detection methods, such as only single feature is exploited for representing highly complicated biomedical images. In this thesis, the second topic is proposing robust multi-stage methods for biomedical image classification and detection.

1.2 Objectives and contributions

Based on the aforementioned motivation, the objectives of this thesis are summarized as follows:

1. to propose a robust classification method with a reject option to systematically select ambiguous and outlier testing samples.
2. to propose a robust biomedical image classification method which can handle the aforementioned challenges of biomedical image variations.

3. to propose a robust biomedical image detection method which can alleviate the limitations of some existing detection methods.

Towards the first object, we investigate Chow's theory and an engineering practice of embedding a support vector machine with a reject option through the Receiver Operating Characteristic (ROC) curve (RO-SVM) [23]. From the study, RO-SVM suffers from the limitation that it can only effectively reject ambiguous samples, i.e., the samples which stay close to the SVM hyperplane but cannot reject the outlier samples. However, outlier samples are supposed to be rejected in Chow's theory as they have low maximum posterior probabilities. In view of this limitation, we propose a twin support vector machine with a reject option (RO-TWSVM) through the ROC curve. The proposed RO-TWSVM enhances the classification robustness through inclusion of an effective rejection rule for ambiguous and outlier samples. RO-TWSVM does not assume that samples being rejected should stay close to the separating hyperplane as RO-SVM does. Since we choose the distance difference in TWSVM to measure the uncertainty of a sample, an outlier sample which is far away from the SVM hyperplane and not rejected by RO-SVM could still be rejected by the proposed RO-TWSVM.

To achieve the second objective, the second study is conducted. Major challenges of biomedical image classification can be concluded as: (i) high level noise; (ii) image blur; (iii) illumination variation; (iv) biomedical patterns with complicated geometric correspondence. These challenges are difficult to be handled simultaneously by using a single classification model trained on images represented based on a single type of local feature. To handle these challenges, we propose a two-stage method for biomedical image classification. At stage 1, we propose a confidence score based classification rule with a reject option for a preliminary decision using the support vector machine (SVM). The testing images going through stage 1 are separated into two groups based on their confidence scores. Those testing images with sufficiently high confidence scores are classified at stage 1 while the others with low confidence scores are rejected and fed to stage 2. At stage 2, the rejected images from stage 1 are first processed by a

subspace analysis technique called eigenfeature regularization and extraction (ERE), and then classified by another SVM trained in the transformed subspace learned by ERE. At both stages, images are represented based on two types of local features, i.e., SIFT and SURF, respectively. They are encoded using various bag-of-words (BoW) models to handle biomedical patterns with and without geometric correspondence, respectively.

Towards the third objective, a study of biomedical image detection is carried out. From the study, it is concluded that supervised detection methods generally outperform unsupervised detection methods. Therefore, in this thesis, we focus on improving supervised methods. Most one-stage supervised detection methods suffer from high computational complexity and “mismatch” effects due to the adoption of sliding windows. Moreover, only global information of each training patch is exploited for training the classifier, which might deteriorate the detection performance. To alleviate the limitations of one-stage supervised methods, we propose a two-stage supervised method for the automated detection of ring-like endosomes in fluorescent microscopy images. The proposed method consists of a localization stage cascaded by an identification stage. Given a testing microscopy image, the localization stage generates a voting-map by locally comparing the selected query patches and the testing image based on a bag-of-words model. Using the voting-map, a number of candidate patches of endosomes are determined. Subsequently, in the identification stage, a support vector machine (SVM) is trained from the training patches represented by a locality-constrained linear coding (LLC) based bag-of-words model. Each of the candidate patches is classified by the SVM to rule out those patches of endosome-like background patterns.

The thesis focuses on proposing robust classification or detection methods to handle uncertain testing samples from both theoretical and engineering perspectives. In Chapter 3, we proposed a twin support vector machine with a reject option (RO-TWSVM) to select ambiguous and outlier samples. The RO-TWSVM is formulated by two stages: training a twin SVM classifier and embedding a reject option using the ROC

curve. In Chapter 4, a two-stage method is proposed for biomedical image classification where the idea of the classification with a reject option is applied to link the two stages of complementary classification methods. In Chapter 5, a two-stage methods for biomedical image detection is proposed where the idea of rejecting outlier samples is exploited in the proposed localization stage to select candidate patches from microscopy images. In general, the three major contributions are related to each other in terms of classification with a reject option and multi-stage structures. The technical details and the related experiments will be described in detail in the corresponding chapters.

1.3 Thesis organization

The remaining parts of this thesis are organized as follows.

Chapter 2 reviews classification with a reject option, twin support vector machine, biomedical image feature extraction, as well as applications in biomedical images with focus on biomedical image classification and detection.

Chapter 3 proposes a novel twin support vector machine with a reject option (RO-TWSVM) through ROC curve. The proposed RO-TWSVM is capable of rejecting both ambiguous and outlier testing samples.

Chapter 4 investigates robust biomedical image classification. The challenges of this application including image variations (e.g., noise, blur, illumination change) and complicated geometric correspondence are analyzed. To handle these challenges, a novel two-stage classification method is proposed.

Chapter 5 investigates robust biomedical image detection. Specifically, a task of detecting ring-like endosome in fluorescent microscopy images is studied. The limitations of one-stage supervised detection methods are analyzed. To alleviate these

limitations, a novel two-stage detection method is proposed.

Chapter 6 concludes the entire thesis and points out several possible future works.

Chapter 2

Literature review

Chapter 2 first reviews classification with a reject option and twin support vector machine in Section 2.1 and 2.2, respectively. Subsequently, biomedical image feature extraction is reviewed in Section 2.3. Finally, applications in biomedical images are reviewed in Section 2.4 with a focus on biomedical image classification and detection.

2.1 Classification with a reject option

In this section, we first review Chow's optimal classification rule with a reject option. Subsequently, we review an engineering practice using Chow's rule, i.e., support vector machine with a reject option (RO-SVM) [23].

2.1.1 Chow's optimal classification rule with a reject option

Classification with a reject option has been studied both theoretically and practically in the literature. Theoretically, Chow [3, 4] developed a Bayesian optimal decision rule with a reject option based on posterior probabilities. Specifically, assuming that $c_1, c_2, \dots, c_M$ denote the categorical labels for an M -categorical classification problem, a testing sample $\mathbf{x}_t$ is:

- rejected whenever

$$\max_{c_i} p(c_i|\mathbf{x}_t) < (1-t), \quad (2.1)$$

- or accepted and labeled to the category c_j whenever

$$\begin{aligned} p(c_j|\mathbf{x}_t) &\geq (1-t), \\ p(c_j|\mathbf{x}_t) &> p(c_i|\mathbf{x}_t) \quad \forall c_i, i \neq j, \end{aligned} \quad (2.2)$$

where $p(c_i|\mathbf{x}_t)$ is the posterior probability of the testing sample $\mathbf{x}_t$ and it can be estimated by certain classification models (see, e.g., [24]).

The parameter t (within $[0, 1]$) controls the trade-off between the error rate and the rejection rate of classification. Specifically, a smaller t leads to a lower error rate and a higher rejection rate. Otherwise, a larger t leads to a higher error rate and a lower rejection rate. If $t = 0$, the error rate is zero and all the samples are rejected. If $t = 1$, the classification rule reduces to the case that no reject option is applied and the error rate reaches the same level as the standard classifier without a reject option [4].

Chow's rule can also be derived under a cost-sensitive framework where classifi-

cation results (e.g., correct classification, misclassification or rejection) are quantified by specific cost values. The optimal model is the one which produces the minimum expected costs. Specifically, assuming that $c_1, c_2, \dots, c_M$ denote the categorical labels for an M -categorical classification problem and c_0 denotes the “rejection” label. The weights w_{ij} ($i = 1, 2, \dots, M$ and $j = 0, 1, 2, \dots, M$) are the costs for classification results. Specifically, w_{ii} denotes the cost for correct classification, w_{ij} ($i \neq j$) denotes the cost for classifying a sample with ground truth label c_i as label c_j , and w_{i0} denotes the cost for rejecting a sample with ground truth label c_i . In this setting, the classification rule with a reject option can be formulated following the notations in [3] using the reject channel function and the decision channel functions. Given a testing sample $\mathbf{x}_t$, the rejection channel function $X_0(\mathbf{x})$ is calculated as:

$$X_0(\mathbf{x}) = \sum_{i=1}^M w_{i0} p(c_i | \mathbf{x}_t). \quad (2.3)$$

The decision channel function $X_j(\mathbf{x}_t)$ ($j = 1, 2, \dots, M$) is calculated as:

$$X_j(\mathbf{x}_t) = \sum_{i=1}^M w_{ij} p(c_i | \mathbf{x}_t). \quad (2.4)$$

The Chow’s rule using the channel functions can be formulated as follows. A testing sample $\mathbf{x}_t$ is:

- rejected whenever

$$X_0(\mathbf{x}_t) < X_j(\mathbf{x}_t) \quad \forall j. \quad (2.5)$$

- or accepted and labeled to the category c_j whenever

$$\begin{aligned} X_j(\mathbf{x}_t) &\leq X_0(\mathbf{x}_t), \\ X_j(\mathbf{x}_t) &< X_i(\mathbf{x}_t) \quad \forall i, i \neq j. \end{aligned} \quad (2.6)$$

It has been proved in [3] that (2.5) and (2.6) are equivalent to (2.1) and (2.2). Note that the first form of Chow’s rule (described in (2.1) and (2.2)) has a parameter t which is

related to the threshold of posterior probabilities while the second form (described in (2.5) and (2.6)) is related to the cost values w_{ij} . An analysis of the relationship between the posterior probability threshold and cost values can be seen in [3].

Based on Chow's rule, some other important statistical theories such as plug-in rules and empirical risk minimization for traditional classification problems have been well generalized to include a reject option [25, 26]. Also, motivated by the theoretical studies, practitioners are seeking for novel strategies to embed a reject option into popular classification models including support vector machines (SVM) [8, 11, 23, 27–29], one-class support vector machines [30], self-organizing map (SOM) [31] and learning vector quantization (LVQ) [32, 33], etc. One of the major problems encountered when exploiting Chow's rule in practical classifiers is that even though the minimum classification error rate is achieved given a fixed (predefined) rejection rate [4], the optimal rejection rate is usually application dependent and difficult to be determined.

For this problem, recent research works focus on embedding reject options from the cost-sensitive point of view [23, 27–29, 34] rather than directly applying posterior probabilities. Among these works, it has been extensively studied to embed Chow's optimal rejection rule into support vector machines (SVM) [35, 36] as SVM is one of the most popular classifiers in practice. Since the original SVM does not output posterior probabilities, Chow's optimal rejection rule [4] cannot be directly applied to SVM. Hence, several methods were proposed to introduce a reject option by manipulating the cost function of SVM [27–29]. However, since these methods constructed the reject option during the training phase, the training procedure had to be re-calculated if the cost settings are changed. Ref. [23] proposed an elegant method to embed a reject option for SVM (RO-SVM) using the Receiver Operating Characteristic (ROC) curve and proved that this method could be applied to any other general binary classifiers [34]. In this thesis, we focus on Tortorella's method to embed a reject option into binary classifiers.

2.1.2 Support vector machine with a reject option using the ROC curve (RO-SVM)

Ref. [34] proposed an optimal rejection rule for binary classification models using the Receiver Operating Characteristic (ROC) curve. Using this method, ref. [23] proposed a support vector machine with a reject option (RO-SVM). In this subsection, we provide a complete review of RO-SVM [23].

In binary classification, support vector machine solves for a hyperplane to discriminate two classes of samples in its original feature space or another feature space through a feature mapping $\phi(\mathbf{x})$ [37]. Such a hyperplane can be formulated as:

$$\mathbf{w}^T \phi(\mathbf{x}) + b = 0, \quad (2.7)$$

where $\mathbf{w}$ is a normal vector to the hyperplane and b is a bias term.

Specifically, given two classes of training samples (here, we simply call one class as “positive class” and the other class as “negative class”) denoted as $\mathbf{x}_n$ and $y_n \in \{-1, +1\}$ ($n = 1, 2, \dots, N$), where N is the number of training samples, $\mathbf{x}_n$ and y_n indicates the feature vector and the categorical label of the n_{th} training sample, respectively. The optimal hyperplane is generally obtained by solving a quadratic programming problem:

$$\begin{aligned} & \underset{\mathbf{w}, b}{\text{minimize}} && \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi_n, \\ & \text{subject to} && y_n(\mathbf{w}^T \phi(\mathbf{x}_n) + b) \geq 1 - \xi_n, \\ & && \xi_n \geq 0, \end{aligned} \quad (2.8)$$

where ξ_n ($n = 1, 2, \dots, N$) is a slack variable introduced for each training sample and a regularization parameter C controls the trade-off between the sum of all the slack variables and the margin width. $\|\cdot\|$ represents the Euclidean norm.

Solving the problem (2.8) by introducing the Lagrange multipliers $a_n \geq 0$ ($n =$

$1, 2, \dots, N$) for each of the constraints in (2.8). The decision function for a new testing data $\mathbf{x}_t$ is determined as:

$$f(\mathbf{x}_t) = \frac{1}{2} \sum_{n=1}^N a_n y_n K(\mathbf{x}_t, \mathbf{x}_n) + b, \quad (2.9)$$

where $K(\mathbf{x}_i, \mathbf{x}_j)$ is the kernel function which is defined as the inner product between the $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$, i.e., $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$. Based on $f(\mathbf{x}_t)$, the classification rule for a testing sample $\mathbf{x}_t$ is:

- classify it as the positive class whenever

$$f(\mathbf{x}_t) \geq 0, \quad (2.10)$$

- or classify it as the negative class whenever

$$f(\mathbf{x}_t) < 0. \quad (2.11)$$

To incorporate a reject option into classification, ref. [23] introduced two parameters t_1 and t_2 to formulate a modified classification rule with a reject option:

- classify it as the positive class whenever

$$f(\mathbf{x}_t) > t_2, \quad (2.12)$$

- or classify it as the negative class whenever

$$f(\mathbf{x}_t) < t_1, \quad (2.13)$$

- or reject it whenever

$$t_1 \leq f(\mathbf{x}_t) \leq t_2, \quad (2.14)$$

where t_1 and t_2 are the thresholds (with $t_1 \leq t_2$) for the value of SVM decision function.

The optimal decision thresholds t_1 and t_2 can be determined under a cost-sensitive framework. Table 2.1 lists the cost settings for different classification results. In this table, P and N denotes the sets of positive and negative class, respectively. CFP , CFN and CR are the costs of false positive errors, false negative errors and rejection, respectively. CTP and CTN are the costs of true positive and true negative, respectively. It is worthwhile to note that the different notations for the cost values are exploited in this section compared with the notation w_{ij} in Section 2.1.1. The reason for such a difference is we focus on binary classification other than a generalized multi-categorical classification problem in this section.

Table 2.1: Cost matrix.

		Predicted Class		
		N	P	Reject
True Class	N	CTN	CFP	CR
	P	CFN	CTP	

Referring to Table 2.1, the expected cost $EC(t_1, t_2)$ can be formulated with thresholds t_1 and t_2 [23]:

$$\begin{aligned}
 EC(t_1, t_2) = & p(P) \cdot CFN \cdot FNR(t_1) + p(N) \cdot CTN \cdot TNR(t_1) \\
 & + p(P) \cdot CTP \cdot TPR(t_2) + p(N) \cdot CFP \cdot FPR(t_2) \\
 & + p(P) \cdot CR \cdot RP(t_1, t_2) + p(N) \cdot CR \cdot RN(t_1, t_2),
 \end{aligned} \tag{2.15}$$

and

$$\begin{aligned}
 TPR(t) &= \int_t^{+\infty} \varphi_+(\omega) d\omega, \quad FNR(t) = \int_{-\infty}^t \varphi_+(\omega) d\omega, \\
 TNR(t) &= \int_{-\infty}^t \varphi_-(\omega) d\omega, \quad RN(t_1, t_2) = \int_{t_1}^{t_2} \varphi_-(\omega) d\omega, \\
 FPR(t) &= \int_t^{+\infty} \varphi_-(\omega) d\omega, \quad RP(t_1, t_2) = \int_{t_1}^{t_2} \varphi_+(\omega) d\omega,
 \end{aligned} \tag{2.16}$$

where FNR , TNR , TPR , FPR are the false negative rate, the true negative rate, the true positive rate and the false positive rate while RP and RN are the rejection rates on the

positive and the negative samples, respectively. $p(P)$ and $p(N)$ are the prior probabilities of the positive and negative classes, respectively. $\varphi_+(\omega) = p(f(\mathbf{x}) = \omega | \mathbf{x} \in P)$ and $\varphi_-(\omega) = p(f(\mathbf{x}) = \omega | \mathbf{x} \in N)$ are class-conditional probability density functions of $f(\mathbf{x})$.

Following the derivations in [23], the expected cost in (2.15) can be rewritten as:

$$EC(t_1, t_2) = \varepsilon_2(t_2) - \varepsilon_1(t_1) + p(P) \cdot CFN + p(N) \cdot CTN, \quad (2.17)$$

where

$$\begin{aligned} \varepsilon_1(t_1) &= p(P) \cdot CFN^* \cdot TPR(t_1) + p(N) \cdot CTN^* \cdot FPR(t_1), \\ \varepsilon_2(t_2) &= p(P) \cdot CTP^* \cdot TPR(t_2) + p(N) \cdot CFP^* \cdot FPR(t_2), \end{aligned} \quad (2.18)$$

and

$$\begin{aligned} CTP^* &= CTP - CR, & CTN^* &= CTN - CR, \\ CFN^* &= CFN - CR, & CFP^* &= CFP - CR. \end{aligned} \quad (2.19)$$

To minimize $EC(t_1, t_2)$, since the term $p(P) \cdot CFN + p(N) \cdot CTN$ does not depend on t_1 and t_2 , minimization of the expected cost $EC(t_1, t_2)$ is equivalent to:

$$\begin{aligned} &\underset{t_1, t_2}{\text{minimize}} && \varepsilon_2(t_2) - \varepsilon_1(t_1), \\ &\text{subject to} && t_1 < t_2. \end{aligned} \quad (2.20)$$

To solve problem (2.20), since $\varepsilon_1(t_1)$ only depends on t_1 and $\varepsilon_2(t_2)$ depends on t_2 , we can maximize $\varepsilon_1(t_1)$ and minimize $\varepsilon_2(t_2)$ simultaneously. The optimization problem can be formulated as:

$$\begin{aligned} &\underset{t_1}{\text{maximize}} && \varepsilon_1(t_1), \\ &\underset{t_2}{\text{minimize}} && \varepsilon_2(t_2), \\ &\text{subject to} && t_1 < t_2. \end{aligned} \quad (2.21)$$

The values of $\varepsilon_1(t_1)$ and $\varepsilon_2(t_2)$ are closely related to TPR and FPR which are the coordinates in the ROC curve. Hence, the optimization problem can be solved geometrically by finding the most appropriate operating points in the ROC curve. Each point in a ROC curve is calculated corresponding to a threshold within the range of $[0, 1]$.

Specifically, t_1 and t_2 are determined by searching the tangential intersection points between the level curves $\varepsilon_1(t_1)$ and $\varepsilon_2(t_2)$ with the slopes (defined by (2.22), (2.23)) and the convex hull of the ROC curve [23] as shown in Fig. 2.1. The two level curve slopes m_1 and m_2 can be calculated based on the level curve (straight line) equations defined in (2.18) as:

$$m_1 = -\frac{p(N) \cdot CTN^*}{p(P) \cdot CFN^*}. \quad (2.22)$$

$$m_2 = -\frac{p(N) \cdot CFP^*}{p(P) \cdot CTP^*}. \quad (2.23)$$

Figure 2.1: The case when condition (2.24) is satisfied and two level curves do not intersect at the same point with the convex hull of the ROC curve.

In [23], a necessary condition (2.24) for the existence of reject option is derived from the constraint $t_1 < t_2$:

$$\frac{CTN \cdot CTP - CFN \cdot CFP}{(CTN + CTP) - (CFN + CFP)} > CR. \quad (2.24)$$

Figure 2.2: The case when condition (2.24) is violated.

Figure 2.3: The case when condition (2.24) is satisfied but two level curves intersect at the same point with the convex hull of the ROC curve.

Fig. 2.2 illustrates the case that condition (2.24) is violated. Condition (2.24) is necessary but not sufficient since two level curves may intersect at the same point in the convex hull of the ROC curve (shown in Fig. 2.3). Therefore, the feasibility of t_1 and t_2 should be checked twice. The first is to confirm if condition (2.24) is satisfied and the second is to check if $t_1 = t_2$ after the geometrical searching procedure.

Once t_1 and t_2 are determined, the classification rule for RO-SVM as shown in Eqs. (2.12) to (2.14) can be applied to select reject or accepted testing samples. From the formulation, it is noted that one of the major limitations of RO-SVM is that it can only effectively reject ambiguous samples, i.e., the samples which stay close to the SVM hyperplane but cannot reject outlier samples. However, outlier samples are supposed to be rejected by Chow's rule as they correspond to low maximum posterior probabilities.

2.2 Twin support vector machine (TWSVM)

Twin Support Vector Machine (TWSVM) [38] combines the spirit of the generalized eigenvalue proximal support vector machine (GEPSVM) [39] with the original SVM optimization formulation. It trains for nonparallel hyperplanes around which the samples of the corresponding class are clustered. The resulted hyperplanes have sparse representations similar to that of SVM's hyperplane [38]. For binary classification, let the positive and negative labeled training samples be represented as the rows of matrices $\mathbf{A}$ and $\mathbf{B}$, respectively [38]. The dimensions of $\mathbf{A}$ and $\mathbf{B}$ are $m_1 \times n$ and $m_2 \times n$, where m_1 and m_2 are respectively the number of positive and negative training samples and n denotes the number of features of a sample. Two nonparallel hyperplanes are parameterized as

$$\begin{aligned} \mathbf{x}^T \mathbf{w}^{(1)} + b^{(1)} &= 0 \\ \mathbf{x}^T \mathbf{w}^{(2)} + b^{(2)} &= 0 \end{aligned} \tag{2.25}$$

where $\mathbf{x}$ denotes the feature vector of a sample, $[\mathbf{w}^{(1)}, b^{(1)}]$ and $[\mathbf{w}^{(2)}, b^{(2)}]$ denote the parameters for the hyperplanes corresponding to the positive and negative classes, respectively.

The two primal quadratic programming problems (QPPs) formulated in [19] are given by

$$\begin{aligned}
 & (TWSVM1) \\
 & \underset{\mathbf{w}^{(1)}, \mathbf{b}^{(1)}, q}{\text{minimize}} \quad \frac{1}{2}(\mathbf{A}\mathbf{w}^{(1)} + \mathbf{1}_1\mathbf{b}^{(1)})^T(\mathbf{A}\mathbf{w}^{(1)} + \mathbf{1}_1\mathbf{b}^{(1)}) + c_1\mathbf{1}_2^T q \\
 & \text{subject to} \quad -(\mathbf{B}\mathbf{w}^{(1)} + \mathbf{1}_2\mathbf{b}^{(1)}) + q \geq \mathbf{1}_2, q \geq 0
 \end{aligned} \tag{2.26}$$

$$\begin{aligned}
 & (TWSVM2) \\
 & \underset{\mathbf{w}^{(2)}, \mathbf{b}^{(2)}, q}{\text{minimize}} \quad \frac{1}{2}(\mathbf{B}\mathbf{w}^{(2)} + \mathbf{1}_2\mathbf{b}^{(2)})^T(\mathbf{B}\mathbf{w}^{(2)} + \mathbf{1}_2\mathbf{b}^{(2)}) + c_2\mathbf{1}_1^T q \\
 & \text{subject to} \quad (\mathbf{A}\mathbf{w}^{(2)} + \mathbf{1}_1\mathbf{b}^{(2)}) + q \geq \mathbf{1}_1, q \geq 0
 \end{aligned} \tag{2.27}$$

where $\mathbf{1}_1$ and $\mathbf{1}_2$ are vectors of ones of appropriate dimensions. c_1 and c_2 are regularization parameters. All inequalities between vectors are componentwise. Essentially, the Twin Support Vector Machine (TWSVM) [38] is proposed to solve the binary classification problem by generating two nonparallel hyperplanes such that each hyperplane is closer to one of the two classes and is as far as possible from the other. The first term of the cost function of (2.26) (similarly for (2.27)) is formulated to minimize the sum of the squared distances from the hyperplane to samples of one class and the constraints require the hyperplane to be at a distance at least 1 from the samples of the other class. The second term of the cost function of (2.26) penalizes the violation of such constraints [38]. By Lagrangian formulation for the primal problems (2.26) and (2.27) and applying Karush-Kuhn-Tucker (K.K.T) conditions, the Wolfe dual problems [40] are obtained as:

$$\begin{aligned}
 & (DTWSVM1) \\
 & \underset{\boldsymbol{\alpha}}{\text{maximize}} \quad \mathbf{1}_2^T \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{G}(\mathbf{H}^T \mathbf{H})^{-1} \mathbf{G}^T \boldsymbol{\alpha} \\
 & \text{subject to} \quad 0 \leq \boldsymbol{\alpha} \leq \mathbf{c}_1
 \end{aligned} \tag{2.28}$$

$$\begin{aligned}
 & (DTWSVM2) \\
 & \underset{\boldsymbol{\gamma}}{\text{maximize}} \quad \mathbf{1}_1^T \boldsymbol{\gamma} - \frac{1}{2} \boldsymbol{\gamma}^T \mathbf{P}(\mathbf{Q}^T \mathbf{Q})^{-1} \mathbf{P}^T \boldsymbol{\gamma} \\
 & \text{subject to} \quad 0 \leq \boldsymbol{\gamma} \leq \mathbf{c}_2
 \end{aligned} \tag{2.29}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are Lagrangian multipliers, $\mathbf{H} = \mathbf{P} = [\mathbf{A} \quad \mathbf{1}_1]$ and $\mathbf{G} = \mathbf{Q} = [\mathbf{B} \quad \mathbf{1}_2]$ and the inequality constraints in (2.28) and (2.29) are for each component of the vectors

α and γ , respectively, i.e., each component of these vectors is between 0 and c_1 (or c_2). When the data are linearly independent, $(\mathbf{H}^T \mathbf{H})^{-1}$ and $(\mathbf{Q}^T \mathbf{Q})^{-1}$ are both positive definite. However, they can be ill-conditioned in some situations [38]. In the original formulation of TWSVM, a regularization term is introduced (i.e., make $\mathbf{H}^T \mathbf{H}$ as $\mathbf{H}^T \mathbf{H} + \varepsilon \mathbf{I}$, $\varepsilon > 0$) to handle the possible ill-conditioning. Solving the optimization problems leads to the nonparallel hyperplanes defined in (2.25). In the testing phase, a testing sample is assigned to the class whose hyperplane is closer to the sample, i.e.,

$$c = \arg \min_{l=1,2} \frac{|\mathbf{x}^T \mathbf{w}^{(l)} + b^{(l)}|}{\|\mathbf{w}^{(l)}\|_2}. \quad (2.30)$$

TWSVM can be extended to a nonlinear classifier by considering surfaces induced by an appropriate kernel function K :

$$\begin{aligned} K(\mathbf{x}^T, \mathbf{C}^T) \mathbf{u}^{(1)} + b^{(1)} &= 0 \\ K(\mathbf{x}^T, \mathbf{C}^T) \mathbf{u}^{(2)} + b^{(2)} &= 0 \end{aligned} \quad (2.31)$$

where $\mathbf{C} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix}$. For any $\mathbf{A} \in \mathbf{R}^{m \times n}$ and $\mathbf{B} \in \mathbf{R}^{n \times k}$, the kernel function $K(\mathbf{A}, \mathbf{B})$ maps $\mathbf{R}^{m \times n} \times \mathbf{R}^{n \times k}$ into $\mathbf{R}^{m \times k}$ [39]. The parameters $\mathbf{u}^{(l)}$ and $b^{(l)}$ ($l = 1, 2$) are determined through quadratic programming. The kernel tricks on TWSVM were well studied in [41]. Fig. 2.4 illustrates the hyperplanes generated by SVM and TWSVM for a 2-dimensional binary classification problem.

Figure 2.4: An illustration to show the differences between the SVM hyperplane and the TWSVM hyperplanes. (a) The SVM hyperplane; (b) The TWSVM hyperplanes.

2.3 Biomedical image feature extraction

In this section, we first review local feature based bag-of-words (BoW) model in Section 2.3.1. Then we briefly review a subspace analysis called eigenfeature regularization and extraction (ERE) in Section 2.3.2.

2.3.1 Local feature based bag-of-words (BoW) model

Recently, local feature based classification or detection methods achieve great success in biomedical classification or detection tasks [42,43]. In this thesis, two proposed methods for biomedical image classification and detection, respectively, are formulated based on representing biomedical images using local feature based bag-of-words model (BoW).

Here, we firstly review three popular local features applied in biomedical image applications, namely scale invariant feature transform (SIFT) [44–46], speeded-up robust features (SURF) [47] and local binary patterns (LBP) [43,48].

SIFT

SIFT, first proposed in [44], describes local gradient information which remains robust across noise, scale and illumination variation within an image. The original proposal of SIFT consists of two steps to extract SIFT descriptors from an image:

1. Detect keypoints from the image through scale-space filtering.
2. Calculate a SIFT descriptor centered at each keypoint.

It is discussed in [44] that Step 1 is implemented mainly for robust image matching. For image classification, Step 1 usually is not implemented and dense SIFT features are more commonly applied, i.e., calculating SIFT descriptors from dense and overlapping grids which cover the whole image [49]. In this thesis, we mainly study and exploit dense SIFT features.

Fig. 2.5 illustrates an example of extracting dense SIFT features from a microscopy image. Given an image, it is divided into $4 \times 2 = 8$ grids. Each grid is then further

Figure 2.5: An example of extracting dense SIFT features from a microscopy image.

divided into $4 \times 4 = 16$ subgrids. For each subgrid, the magnitude and orientation of the gradient for each pixel are calculated. The black arrows denote the gradients where the length and direction of each arrow are the magnitude and the orientation of the gradient, respectively. Finally, A histogram of gradients with respect to 8 orientation bins (in the range of 0 to 360 degrees) is built to summarize the gradients within each subgrid. A gradient is assigned to its nearest orientation bin by comparing the gradient's orientation with the 8 orientation bins. The weight of the assignment is the magnitude of the gradient. If a gradient has its orientation in the middle of 2 neighboring orientation bins, a bilinear interpolation is applied. By this procedure, each grid shown in Fig.2.5(a) can be represented by concatenating 16 8-bin histograms, i.e., a $16 \times 8 = 128$ dimensional SIFT feature vector (descriptor). In this example, the whole image can be decomposed into 8 128-dimensional SIFT descriptors.

SURF

SURF [47] was proposed as a fast calculated local feature inspired by SIFT. Similar to SIFT, the original proposal of SURF [47] consists of two steps, i.e., keypoint detection and descriptor calculation, but different techniques are applied:

1. Detect keypoints from the image using blob detectors based on Hessian matrix.
2. Calculate a SURF descriptor from a region centered at each keypoint based on the integral image.

Similar to dense SIFT, we can also extract SURF descriptors from densely overlapped grids within an image, i.e., dense SURF. Each grid is split into 4×4 (pixels) subgrids. From each subgrid, the Haar wavelet responses are extracted at 5×5 (pixels) regularly spaced sample points. The responses are weighted with a Gaussian to offer more robustness for deformations, noise, and translation [47].

SURF can be computed much faster than SIFT due to its adoption of integral image [47]. Moreover, it has been discussed in [50, 51] that SURF and SIFT enjoy complementary merits. Specifically, SIFT is more robust to changes in scale, image blur and rotation while SURF is more robust to changes in illumination.

LBP

Local binary pattern (LBP) [48] is a powerful local texture feature descriptor. LBP achieves impressive performance in applications including face recognition, texture classification, human detection, biomedical image classification, etc. The basic procedure to generate LBP features is reviewed as follows.

Given an image, for each pixel, its LBP descriptor can be calculated in the neighborhood centered at the pixel as:

$$LBP_{P,R} = \sum_{i=0}^{P-1} s(g_i^R - T)2^i, \quad s(x) = \begin{cases} 1, & x \geq 0. \\ 0, & x < 0. \end{cases} \quad (2.32)$$

where P is the number of selected pixels in the neighborhood, R is the radius (in pixels) from the center pixel to the local neighboring pixels, g_i^R denotes the i_{th} pixel value in the neighborhood which is R pixels away from the center pixel and T is the threshold. Different pairs of (P, R) determine different LBP configurations. The most typical settings are (8,1), (12, 2) and (16, 4) [48].

LBP is a discriminative local feature for biomedical image representation as it has been proved to be robust against biomedical image noise, illumination changes etc. Ref. [43] compared recent LBP-based methods in the biomedical classification tasks using

four benchmark datasets. These LBP variants are listed as follows [43]:

- LBP-r, the standard LBP with rotation invariant.
- EBP, the elliptical LBP.
- LBP_{su2}, scale-adaptive and subuniform-based rotation invariant LBP.
- LBP_{nr}, scale-adaptive LBP without rotation invariant.
- LBP-HF, LBP histogram Fourier features.
- LDBP, the local difference magnitude binary pattern invariant.
- CLBP, in which three types of features are combined to form the completed LBP features.
- CoALBP, in which the co-occurrence among adjacent LBPs have been considered.

Moreover, [43] proposed a novel LBP variant called spatial adjacent histogram based on adaptive local binary patterns (SAHLBP) for biomedical image classification. SAHLBP modifies the original LBP by assigning an adaptive neighborhood radius for each pixel and proposes a spatial adjacent histogram strategy to encode micro-structures for biomedical images. It has been discussed in [43] on three coding schemes (based on different selections of the value of threshold T in (2.32)) conducted for local binary pattern generation. Correspondingly, there are three types of SAHLBP, namely SAHLBP T_i ($i = 1, 2, 3$). The SAHLBP features combined with SVM outperforms the state-of-the-art LBP based methods as listed above. In this thesis, SAHLBP is one of the most important state-of-the-art methods which our method is compared with in Chapter 4.

The above reviewed local features capture distinct information from small regions within images. In this way, an image can be decomposed into multiple local features. Image classification and detection tasks require a global and discriminative representation of images, i.e., techniques to summarize local features to be a discriminative vector

associated with the whole image. One simple solution to building image representation is directly calculating histograms of local features, e.g., LBP histograms [43, 48]. Such direct histogram calculation can be applied to those local features with a limit number of potential feature descriptors, e.g., LBPs. However, local features like SIFT or SURF have an unlimited number of potential descriptors and therefore a direct calculation of histograms cannot be applied. Inspired by text categorization, [52] proposed bag-of-words (BoW) to build image representation by a histogram of a limited number of representative local features. BoW model is popular and successful in the fields of image classification [49, 53], retrieval [52] and detection [54]. The procedure to generate a BoW model is reviewed as follows.

Firstly, local features, e.g., dense SIFT, are extracted from a set of training images. In this way, each training image is decomposed into a number of local features. Then, these local features are clustered into a number of classes using unsupervised methods, e.g., K-means clustering. The centroids of these classes are defined as the visual words and the set of the visual words is defined as the dictionary. By vector quantizing each local feature extracted from an image with respect to the visual words, a histogram which indicates the frequencies of the local features being assigned to each visual word is generated. Such histogram can be regarded as a feature vector for an image. Fig. 2.6 illustrates dense SIFT feature based BoW model and image representation using the BoW model.

To improve the performance of the original BoW, two modification techniques are proposed. The first is spatial pyramid matching (SPM) [49] which is applied to alleviate loss of spatial information by introducing moderate spatial constraints. SPM divides an image into different levels. For example, if the number of levels of SPM is set at 3, the original image is uniformly divided into 1×1 (the whole image), 2×2 and 4×4 regions. A BoW histogram is generated from each region and the final image representation is the concatenated vector of all the BoW histograms.

The second technique is locally-constrained linear coding (LLC) [53]. Instead of

Figure 2.6: (a) The training images are decomposed into many SIFT features. By K-means clustering, these SIFT features are clustered to a predefined number of classes. The centroids of these clusters are referred to as the visual words. In this example, three visual words are generated (shown by black crosses). The set containing all the visual words is referred to as a dictionary. (b) Given an image to be represented by a bag-of-words model, it is first decomposed into several SIFT features. Each SIFT feature is assigned to its nearest visual word using Euclidean distance. Implementing such feature assignment for each SIFT feature, the image can be represented as a histogram of visual word frequencies, i.e., the counts of the occurrences of the SIFT features assigned to each visual word.

selecting the nearest visual word, for a feature $\mathbf{f}_i$, LLC firstly selects K nearest neighbors of $\mathbf{f}_i$ in the dictionary to form a local base $\mathbf{B}_i = [\mathbf{b}_{1i}, \mathbf{b}_{2i}, \dots, \mathbf{b}_{Ki}]$. Then it solves the following linear system to obtain the code $\mathbf{c}_i = [c_{i1}, c_{i2}, \dots, c_{iK}]^T$ for $\mathbf{f}_i$:

$$\begin{aligned} \min_{\mathbf{c}_i} & \|\mathbf{f}_i - \mathbf{B}_i \mathbf{c}_i\|^2, \\ \text{s.t. } & \mathbf{1}^T \mathbf{c}_i = 1. \end{aligned} \quad (2.33)$$

The constraint of the linear system guarantees the summation of the entries of the code $\mathbf{c}_i$ is 1. LLC generates a sparse and discriminative image representation which can be classified by linear classifiers [53].

2.3.2 Eigenfeature regularization and extraction

To obtain more discriminative representations of images, subspace analysis can be conducted for feature dimensionality reduction. For example, principal component analysis (PCA) is a widely used subspace analysis technique which is defined as the orthogonal projection of data onto lower dimensional linear space (principal subspace), where the variance of the projected data is maximized [55]. Subspace analysis is widely used for feature dimensionality reduction, lossy data compression and data visualization [56]. A good review of subspace analysis on image classification can be seen in [57].

Here, we provide a brief review of a subspace analysis method called eigenfeature regularization and extraction (ERE) [58] as this technique is applied (in Chapter 4) to find discriminative subspace to classify biomedical images from a small number of training images. ERE is highly effective in learning discriminative subspaces especially for those applications where a large number of training samples is hard to be obtained.

The goal of ERE [58] is to extract discriminative structures underlying a small number of training samples. In general, ERE applies eigenfeature regularization to replace large eigenvalues of the within-class scatter matrix with designed values. Then a discriminative subspace is obtained by eigen decomposition on the regularized total scatter matrix. The image representations in this subspace are expected to be classified

with high accuracy.

Specifically, given a training dataset with M categories and q_m denotes the number of images for the m_{th} category. Each training image is represented by an l -dimensional feature vector $\mathbf{x}_{mn}$, where $\mathbf{x}_{mn}$ represents the n_{th} sample for the m_{th} category. Assuming the prior probability of each category is the same, i.e., $1/M$, the within-class scatter matrix $\mathbf{S}^w$ can be calculated as:

$$\mathbf{S}^w = \frac{1}{M} \sum_{m=1}^M \frac{1}{q_m} \sum_{n=1}^{q_m} (\mathbf{x}_{mn} - \bar{\mathbf{x}}_m)(\mathbf{x}_{mn} - \bar{\mathbf{x}}_m)^T, \quad (2.34)$$

where $\bar{\mathbf{x}}_m = \frac{1}{q_m} \sum_{n=1}^{q_m} \mathbf{x}_{mn}$. $\mathbf{S}^w$ can be decorrelated by eigen decomposition as:

$$\mathbf{\Lambda}^w = \mathbf{\Phi}^{wT} \mathbf{S}^w \mathbf{\Phi}^w, \quad (2.35)$$

where $\mathbf{\Phi}^w = [\boldsymbol{\phi}_1^w, \boldsymbol{\phi}_2^w, \dots, \boldsymbol{\phi}_l^w]$ is the eigenvector matrix of $\mathbf{S}^w$ and $\mathbf{\Lambda}^w$ is the diagonal matrix of eigenvalues $\lambda_1^w, \dots, \lambda_l^w$ associated with the eigenvectors. In this context, the eigenvalues are sorted in a descending order, i.e., $\lambda_1^w \geq \dots \geq \lambda_l^w$. It is discussed in [58] that the large eigenvalues are more reliable and the small eigenvalues correspond to noise dominant subspaces. The beginning of the unreliable eigenvalues (indexed by $m+1$) is estimated by [58]:

$$\lambda_{m+1}^w = \max\{\forall \lambda_k^w | \lambda_k^w < (\lambda_{med}^w + \mu(\lambda_{med}^w - \lambda_r^w))\}, \quad (2.36)$$

where $\lambda_{med}^w = \text{median}\{\forall \lambda_k^w | k \leq r\}$, r is the rank of $\mathbf{S}^w$ and μ is a constant. The value of μ is set at 1 as suggested in [58].

To handle the lack of training samples, ERE assumes an eigenspectra model underlying the obtained within-class scatter matrix $\mathbf{S}^w$ and this model consists of two parameters α and β calculated by [58]:

$$\alpha = \frac{\lambda_1^w \lambda_m^w (m-1)}{\lambda_1^w - \lambda_m^w}, \quad (2.37)$$

$$\beta = \frac{m\lambda_m^w - \lambda_1^w}{\lambda_1^w - \lambda_m^w}. \quad (2.38)$$

The original eigenvalues of $\mathbf{S}^w$ are replaced by the regularized version as follows [58]:

$$\tilde{\lambda}_k^w = \begin{cases} \lambda_k^w & k < m, \\ \frac{\alpha}{k+\beta} & m \leq k \leq r, \\ \frac{\alpha}{r+1+\beta} & r < k \leq l. \end{cases} \quad (2.39)$$

The corresponding eigenvector matrix is replaced by:

$$\tilde{\Phi}^w = \left[\frac{\phi_1^w}{\sqrt{\tilde{\lambda}_1^w}}, \dots, \frac{\phi_l^w}{\sqrt{\tilde{\lambda}_l^w}} \right]. \quad (2.40)$$

Based on the regularized eigenvector matrix, an original sample $\mathbf{x}_{mn}$ is transformed by:

$$\tilde{\mathbf{y}}_{mn} = \tilde{\Phi}^{wT} \mathbf{x}_{mn}. \quad (2.41)$$

Consequently, a regularized total scatter matrix $\tilde{\mathbf{S}}^t$ is calculated as:

$$\tilde{\mathbf{S}}^t = \frac{1}{M} \sum_{m=1}^M \frac{1}{q_m} \sum_{n=1}^{q_m} (\tilde{\mathbf{y}}_{mn} - \bar{\mathbf{y}})(\tilde{\mathbf{y}}_{mn} - \bar{\mathbf{y}})^T, \quad (2.42)$$

where $\bar{\mathbf{y}} = \frac{1}{M} \sum_{m=1}^M \frac{1}{q_m} \sum_{n=1}^{q_m} \tilde{\mathbf{y}}_{mn}$. The eigen decomposition is applied on $\tilde{\mathbf{S}}^t$. The d largest eigenvalues are selected and correspondingly an eigenvector matrix $\tilde{\Phi}_d^t$ for dimensionality reduction is obtained as:

$$\tilde{\Phi}_d^t = [\tilde{\phi}_1^t, \dots, \tilde{\phi}_d^t]. \quad (2.43)$$

In general, a sample $\mathbf{x}$ in the original vector space is regularized by $\tilde{\Phi}^w$ and transformed to the subspace as [58]:

$$\mathbf{f} = (\tilde{\Phi}^w \tilde{\Phi}_d^t)^T \mathbf{x}. \quad (2.44)$$

For brevity, let $\mathbf{U} = \tilde{\Phi}^w \tilde{\Phi}_d^t$ and $\mathbf{f} = \mathbf{U}^T \mathbf{x}$ represents the transformed feature. The value

of d is determined based on the grouping boundary of the eigenvalues [58].

2.4 Applications in biomedical images

In this section, we briefly review the applications in biomedical image classification and detection.

2.4.1 Biomedical image classification

Biomedical images are playing significant roles in modern biological studies and medical diagnosis processes. Classification of different types of images is a critical step for biomedical applications including protein structure and function understanding [13], pre-cancer predictions [59] and autoimmune disease detection [60]. Traditionally, biomedical images are visually classified by human experts. Hence, the results of classification highly depend on human expertise and therefore they are prone to subjectivity. Manual classification also suffers from inefficiency due to high-throughput production of images in practice. Therefore, it is greatly demanded by the biomedical research community for automating the process of image classification to reduce human burden. To alleviate the problems encountered by manual classification, many research works on biomedical image classification develop machine learning methods to automate the process (see e.g., [12, 22]). For example, pioneering works have been carried out for identification of subcellular localization of proteins based on microscopic image classification [13, 17, 18, 61].

In this section, we briefly review several local feature based classification methods for biomedical images. The basic procedure of local feature based methods consists of two steps: feature extraction and classifier training.

- **Feature extraction:** Discriminative local features are first extracted from training images. Then, the training images are represented by these local features.

- **Classifier training:** With the training image representations, a classifier is trained to discriminate between biomedical images belonging to different classes.

For image representation, local features, such as scale invariant feature transform (SIFT) [44–46], speed-up robust features (SURF) [42, 47] and local binary patterns (LBP) [43, 48] are shown to be effective in biomedical image classification as these local features and their variants can capture discriminative local micro-structures of biomedical patterns. Based on these local features, image representations such as the BoW model are built for biomedical images. The details of these local features and the BoW model have been reviewed in Section 2.3. After building up discriminative image representations, classification models, such as support vector machines [62, 63] and artificial neural networks [64, 65] are trained on images represented by their extracted feature vectors. The trained model can be applied to automatically classify unseen testing images.

Currently, automated biomedical image classification remains a challenging problem because image quality is highly influenced by high level noise, image blur and illumination variation etc. Moreover, biomedical patterns have very complicated geometric correspondence properties. It is observed that some categorical biomedical patterns have high geometric correspondence, e.g., they show the geometrical shapes, edges, boundaries and centroids. However, some other categorical patterns might have low or no geometric correspondence, e.g., they have similar texture features but do not have consistent shapes. These challenges are difficult to be handled simultaneously by using a single classification model trained on images represented based on a single type of local features.

2.4.2 Biomedical image detection

The second application in biomedical images studied in this thesis is biomedical image detection, i.e., determining locations of a specific type of biological structures in relatively a big image. Compared with biomedical image classification, the goal of

biomedical image detection is not only to predict the identity (or categorical label) of an image, but also to determine the locations of the patterns of biological interest.

Several methods for detecting subcellular structures in fluorescent microscopy images have been proposed in the literature. Ref. [66] carried out a quantitative evaluation of multiple methods for spot detection. In general, detection methods can be divided into two groups: unsupervised and supervised methods.

Unsupervised detection methods

Unsupervised methods require some prior knowledge of the appearance model of the target to be detected and usually contain parameters to be tuned either manually or semi-automatically [66]. The prior information is applied for carefully designing detectors based on image processing techniques such as local background subtraction [67], linear or morphological image filtering [68] and wavelet-based multiscale detectors [69] etc. For instance, ref. [70] applied the fast radial symmetry transformation (FRST) [71] to detect cell nuclei based on the information that cell nuclei are approximately circular and symmetrical. Since we choose this method as a representative of unsupervised methods to compare with our proposed method, we provide a review of this method as follows.

To detect the centroids of nuclei within the radius range as $\rho_{\min} \leq \rho \leq \rho_{\max}$ (pixels), at first, the gradient $g(x)$ is calculated at each pixel x . Subsequently, for each radius value ρ , a positively affected pixel $P_+(x)$ and a negatively affected pixel $P_-(x)$ are determined. The positively affected pixel $P_+(x)$ is the pixel which the gradient $g(x)$ is pointing to, at a distance ρ from the location of the pixel x . The negatively affected pixel $P_-(x)$ is the pixel which the gradient $g(x)$ is pointing away from, at a distance ρ from the location of the pixel x . Specifically, $P_+(x)$ and $P_-(x)$ can be calculated as:

$$P_+(x) = x + \text{Round} \left(\frac{g(x)}{\|g(x)\|} \rho \right). \quad (2.45)$$

$$P_-(x) = x - \text{Round} \left(\frac{g(x)}{\|g(x)\|} \rho \right). \quad (2.46)$$

Then an orientation projection image O_ρ and a magnitude projection image M_ρ are initially zero. Based on $P_+(x)$ and $P_-(x)$, the corresponding pixel $P_+(x)$ in the image O_ρ and in the image M_ρ is incremented by 1 and $\|g(x)\|$, respectively, while the corresponding pixel $P_-(x)$ is decremented by the same quantities.

Based on O_ρ and M_ρ , a radial symmetry image at radius ρ is calculated by the convolution as:

$$S_\rho = F_\rho * G_\rho, \quad (2.47)$$

where

$$F_\rho = \frac{M_\rho(x)}{k_\rho} \left(\frac{|\tilde{O}_\rho(x)|}{k_\rho} \right)^\gamma. \quad (2.48)$$

$$\tilde{O}_\rho(x) = \begin{cases} O_\rho(x) & \text{if } |O_\rho(x)| < k, \\ k_\rho & \text{otherwise.} \end{cases} \quad (2.49)$$

G_ρ is a Gaussian filter, γ is the radial strictness parameter and k_ρ is a scaling factor.

The total FRST transform of the image is defined as the average of the symmetry contributions over all radii as:

$$S = \frac{1}{N} \sum_{\rho \in N} S_\rho, \quad (2.50)$$

where N is the set of the potential radius. The detection of nuclei is implemented by thresholding the FRST transformed image S followed by a sequence of morphological operations [70]. Hence, we call this method MO-FRST in the experimental section in Chapter 5.

Supervised detection methods

Supervised methods first “learn” the discriminative appearance information of the target from annotated patches, i.e., small regions of interest (ROI) which contain an isolated target (positive patches) or an irrelevant background pattern (negative patches). Then it searches in the big testing images for the patches of targets from a classification perspective. Supervised methods require a training phase before testing. In the training

phase, a classifier is trained to discriminate between positive and negative patches. In the testing phase, given a big microscopy image, sliding windows, i.e., overlapping small patches which scan over a relatively big image, are adopted to search all the locations in the microscopy image. Each window is fed to the pretrained classifier to determine whether it contains the target or not. In this thesis, the aforementioned sliding window based supervised methods are called as one-stage methods to distinguish them from our proposed two-stage supervised method in Chapter 5. We mainly compared our method with two supervised methods reviewed in [66] as follows:

- AB-Haar method (also called Viola-Jones detectors [72]), i.e., the Adaboost classifier trained on Haar-like features. This method was modified for biomedical image detection in [73].
- LDA-Pixel method, which as proposed in [66] for spot detection from microscopy images.

In general, most unsupervised methods are computationally efficient and perform well when the targets have simple and distinct features, e.g., detecting circular targets from highly non-circular background patterns. However, unsupervised methods normally fail to detect complicated targets especially when the adopted prior information is not sufficiently discriminative to distinguish the targets from some background patterns. In such scenarios, supervised methods generally outperform unsupervised methods as they apply the discriminative information between the targets and the non-target background patterns. Therefore, in this thesis, we focus on improving supervised methods.

A disadvantage of one-stage supervised methods is the adoption of the sliding window search as it is generally computationally complicated and might miss the detection if the target is partially covered or “mismatched” with the sliding window, e.g., a small target stays inside a relatively large sliding window. Also, the most important factor for a one-stage supervised method is the performance of its classifier.

Such performance is highly dependent on the selection of training patches as well as whether each training patch is represented by a discriminative feature vector. Most one-stage supervised methods only exploit global information of each training patch, i.e., representing a training patch (as a whole) by a feature vector, which might deteriorate the detection performance.

Chapter 3

Twin SVM with a reject option through ROC curve (RO-TWSVM)

Chapter 3 proposes a twin support vector machine with a reject option (RO-TWSVM) for binary classification. Section 3.1 analyzes the limitation of RO-SVM and the merits of TWSVM. In Section 3.2, we present the formulation of RO-TWSVM. In Section 3.3, the performance of RO-TWSVM is evaluated on two synthetic datasets. We conduct three statistical tests using multiple real-world benchmark datasets to compare RO-TWSVM with the competing methods in Section 3.4. Section 3.5 summarizes this chapter.

3.1 Introduction

As reviewed in Section 2.1.2, one of the major limitations of RO-SVM [23] is that it can only effectively reject ambiguous samples, i.e., the samples which stay close to the SVM hyperplane, but cannot reject outlier samples. As discussed in Section 2.2, twin support vector machine (TWSVM) is a novel SVM variant. Comparing with the SVM, TWSVM is computationally more efficient and has better classification performance for imbalanced data [38]. This is due to the flexibility offered by the nonparallel hyperplanes trained with TWSVM.

Motivated by these observations, we propose a twin support vector machine with a reject option (RO-TWSVM) through the Receiver Operating Characteristic (ROC) curve. The proposed RO-TWSVM enhances the classification robustness through inclusion of an effective rejection rule for ambiguous and outlier samples. RO-TWSVM does not assume that samples being rejected should stay close to the separating hyperplane as RO-SVM does. Since we choose the distance difference in TWSVM to measure the uncertainty of a sample, an outlier sample which is far away from the SVM hyperplane and not rejected by RO-SVM could still be rejected by the proposed RO-TWSVM.

Extensive experiments are conducted on multiple synthetic and real-world datasets using multiple statistical significance tests. Significant performance improvement of the proposed RO-TWSVM over TWSVM-without-RO is observed in view of classification costs. Also, for the datasets which can be easily discriminated by one hyperplane, the two methods achieve comparable performance. Whereas, RO-TWSVM significantly outperforms RO-SVM for datasets which are difficult to be discriminated by one hyperplane.

3.2 Methodology

Essentially, TWSVM encodes discriminative information through distances from samples to different hyperplanes of the corresponding classes. For binary classification, the decision rule is based on such distance difference:

$$d(\mathbf{x}) = d_-(\mathbf{x}) - d_+(\mathbf{x}), \quad (3.1)$$

where $d_+(\mathbf{x})$ and $d_-(\mathbf{x})$ are the distances from the sample $\mathbf{x}$ to the positive and the negative hyperplanes, respectively. They are defined as:

$$d_+(\mathbf{x}) = \frac{|\mathbf{x}^T \mathbf{w}^{(1)} + b^{(1)}|}{\|\mathbf{w}^{(1)}\|_2}, \quad (3.2)$$

$$d_-(\mathbf{x}) = \frac{|\mathbf{x}^T \mathbf{w}^{(2)} + b^{(2)}|}{\|\mathbf{w}^{(2)}\|_2}, \quad (3.3)$$

where $\|\cdot\|_2$ is the Euclidean norm and $\mathbf{w}^{(l)}, b^{(l)}, l = 1, 2$, are defined in (2.25).

Based on $d(\mathbf{x})$, the classification rule of TWSVM [38] is:

$$\begin{cases} d(\mathbf{x}) \geq 0 & \text{positive class,} \\ d(\mathbf{x}) < 0 & \text{negative class.} \end{cases} \quad (3.4)$$

Inspired by RO-SVM [23] reviewed in Section 2.1.1, we propose a reject region for TWSVM for binary classification delimited by two thresholds t_1 and t_2 ($t_2 > t_1$) based on $d(\mathbf{x})$. Hence, a new classification rule embedding with a reject option can be formulated as:

$$\begin{cases} d(\mathbf{x}) > t_2 & \text{positive class,} \\ d(\mathbf{x}) < t_1 & \text{negative class,} \\ t_1 \leq d(\mathbf{x}) \leq t_2 & \text{rejection.} \end{cases} \quad (3.5)$$

This formulation is reasonable since the reject region is defined as the region where samples have similar distances to both hyperplanes. Fig. 3.1 gives an illustration of a simple 2D example, where two black dashed-dotted lines represent the hyperplanes

Figure 3.1: The reject region for a 2D example. The blue dashed lines delimit the rejection boundaries and the black dashed-dotted lines represent nonparallel hyperplanes trained by TWSVM.

(lines) trained by TWSVM. The region delimited by two blue dashed lines is the reject region and the black diamond represents a sample being rejected. The illustration shows an example of an ambiguous sample. It is noted that an outlier sample could be rejected by this formulation as it could produce a small $d(\mathbf{x})$.

The optimal thresholds t_1 and t_2 can be determined following the principle of minimization of expected cost and using the ROC based method as reviewed in Section 2.1.2 for a specific cost setting. Compared with RO-SVM, the output of SVM decision function $f(\mathbf{x})$ is replaced by the distance difference $d(\mathbf{x})$ of TWSVM defined in (3.1). Thereby, the class-conditional probability density functions are redefined as $\varphi_+(\omega) = p(d(\mathbf{x}) = \omega | \mathbf{x} \in P)$ and $\varphi_-(\omega) = p(d(\mathbf{x}) = \omega | \mathbf{x} \in N)$, respectively.

Algorithm 1 summarizes the procedure of RO-TWSVM. In general, the algorithm consists of two stages. First, a standard TWSVM model is trained over a set of training

data. Subsequently, a reject option is incorporated by determining the values of the two thresholds on the distance difference.

In the first stage, We train a TWSVM model for classification without embedding a reject option. Therefore, standard cross validation is adopted for kernel selection and parameter tuning. In this work, we consider the linear kernel and radius basis function (RBF) kernel. For both the linear kernel and the RBF kernel, the parameters c_1 and c_2 of TWSVM are obtained through a search in the range of $[2^{-8}, 2^8]$. For the RBF kernel, the kernel width σ is obtained through a search in the range of $[2^{-8}, 2^8]$. These parameters are determined through the 5-fold cross validation (see details in [38, 41]).

In the second stage, we embed a reject option into the TWSVM by selecting the proper values of t_1 and t_2 based on the cost settings. We first check if condition (2.24) is satisfied, otherwise, the reject option is not activated and we record this case as “RO is not applicable”. If condition (2.24) is satisfied, we find the tangential intersection points between the level curves and the convex hull of the ROC curve. From these intersection points we obtain the corresponding thresholds t_1 and t_2 . Subsequently, we check if the

Algorithm 1: The description of RO-TWSVM.

Input:
 Data: $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$;
 Costs: CTN, CFP, CFN, CTP, CR ;
Output:
 Reject region: $[t_1, t_2]$ or Message: “RO is not applicable”;

```

1 if condition (2.24) is not satisfied then
2   return Message: “RO is not applicable”;
3 else
4   Randomly divide  $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$  into a training set  $P_k$  and a validation set  $P_{\bar{k}}$ ,
   respectively.
5   Use the samples in the training set  $P_k$  to train for a TWSVM model:  $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, b^{(1)}, b^{(2)}$ .
6   Use the samples in the validation set  $P_{\bar{k}}$  to generate the ROC curve.
7   Determine  $t_1$  and  $t_2$  geometrically through the ROC curve;
8   if  $t_1 < t_2$  then
9     return Reject region:  $[t_1, t_2]$ ;
10  else
11    return Message: “RO is not applicable”;

```

constraint $t_1 < t_2$ is satisfied. If it is satisfied, the reject region is determined and the rule defined in (3.5) is applied. Otherwise, the case is also recorded as “RO is not applicable”.

Remark. The ROC based method to determine the optimal thresholds t_1 and t_2 for RO-TWSVM is similar to RO-SVM [23]. On the one hand, some advantages of RO-SVM are retained in the proposed RO-TWSVM, such as the training for TWSVM hyperplanes and the plot of ROC curve are implemented only once since changes made on cost settings, if any, only influence the slopes of the level curves. Hence, excessive re-calculations for the TWSVM hyperplanes and the ROC curve are not required. On the other hand, RO-TWSVM does not assume that samples being rejected should stay close to the separating hyperplane as RO-SVM does. Since we choose the distance difference in TWSVM to measure the uncertainty of a sample, an outlier which is far away from the SVM hyperplane and not rejected by RO-SVM could still be rejected by the proposed RO-TWSVM.

The effectiveness and advantages of RO-TWSVM will be demonstrated through the experiments on both synthetic and real-world benchmark datasets.

3.3 Experiments on synthetic datasets

In this section, two synthetic datasets are simulated to observe the behavior of the proposed method. The simulation setup here is similar to that in RO-SVM [23] on synthetic dataset 1 [23] to observe the variation of the reject region generated by the proposed RO-TWSVM with respect to various cost settings. TWSVM formulates a hyperplane for samples of each class where better flexibility than that of SVM could be anticipated. Hence, the goal of the second synthetic dataset is to illustrate the merit of this flexibility leading to more reasonable reject regions by RO-TWSVM than those by RO-SVM. In our experiments, the prior probabilities for the two classes are assumed to be equal. Hence, 50% of samples are randomly chosen for training and validation in Procedure 1. The remaining 50% of the samples are used for testing. Among the training

and validating samples, for the RO-TWSVM formulation described in Algorithm 1, the training set P_k contains 75% of the samples and the validation set P_k contains the remaining 25% for determining the optimal thresholds t_1 and t_2 .

3.3.1 Synthetic data 1: uniformly distributed data with Gaussian noise

To demonstrate that the reject regions of RO-TWSVM can adapt to different cost settings, the proposed RO-TWSVM is implemented on a synthetic dataset according to [23]. As shown in Fig. 3.2a, this dataset consists of 720 samples being equally distributed between the two classes. The samples are uniformly distributed with unit variance and spherical Gaussian noise along two $2\pi/3$ arcs with radii 6.2 and 10.0 for the positive and negative classes, respectively.

Table 3.1: The cost models for synthetic data 1.

	CTP	CFP	CTN	CFN	CR
C1	-20	22	-10	18	4
C2	-20	20	-10	18	2
C3	-14	22	-10	24	2
C4	-8	26	-10	28	2
C5	-4	36	-6	120	2
C6	-6	120	-4	36	2

The purpose of this example is to show the variation of reject regions generated by the RO-TWSVM under different cost models. An RBF kernel is adopted and the cost settings applied are shown in Table 3.1. These settings follow those recommended in [23] for RO-SVM where P and N denote the sets of positive and negative classes, respectively. CFP , CFN and CR are the costs of false positive errors, false negative errors and rejection, respectively. CTP and CTN are the costs of true positive and true negative, respectively. An illustration with several reject regions corresponding to the cost settings in Table 3.1 is presented in Fig. 3.2b. In Fig. 3.2b, there is no reject region generated for cost model C1 and C2. The reasons are that C1 violates condition

(a)

(b)

Figure 3.2: (a) Synthetic dataset 1, (b) The reject region boundaries of RO-TWSVM for each cost model: C3 (black solid line), C4 (blue dashed line), C5 (cyan dotted line) and C6 (magenta dashed-dotted line).

(2.24) and the two level curves intersect at the same point for C2. From Table 3.1, it is observed that C3 and C4 represent the cases that $CTP \approx CTN$ and $CFP \approx CFN$

which correspond to applications with balanced costs. Consequently, for these two cost models, the generated reject regions show no rejection preference to either class. In contrary, C5 and C6 represent applications with imbalanced cost settings. In particular, C5 is for the case where the cost of false negative error is much higher than that of false positive error. As a result, we observe that the reject region for C5 includes more samples from the positive class. However, C6 has an opposite setting where the cost of false positive error is much higher than that of false negative error. In this case, the reject region includes more samples from the negative class. We can see both the sizes and locations of reject regions adapt to changes of cost settings. It should be emphasized that this dynamic behavior is desirable since it reflects that our proposed RO-TWSVM can output an appropriate reject region according to a given cost setting.

Table 3.2: The expected costs for different cost models for synthetic data 1.

	C3	C4	C5	C6
RO-TWSVM	-7.5125	-4.5750	-1.3667	-0.2750
RO-SVM	-7.4750	-4.5292	-1.2458	-0.2875

For synthetic dataset 1, our proposed RO-TWSVM generates a reject region similar to that of RO-SVM. To establish a quantitative comparison, we calculate the expected costs of RO-TWSVM and RO-SVM for C3 through C6. The results of these costs are shown in Table 3.2. From Table 3.2, it is observed that RO-TWSVM and RO-SVM produce almost the same costs for the four cost models. Therefore, for this dataset, our proposed RO-TWSVM achieves comparable performance with RO-SVM. In Fig. 3.3a and Fig. 3.3b, we plot the class-conditional probability density functions (pdfs) of $d(\mathbf{x})$ (defined in (3.1)) for TWSVM and that of SVM outputs [23], respectively. The overlapping regions of class-conditional pdfs of $d(\mathbf{x})$ reflect those samples which have similar distances to both hyperplanes of TWSVM (see Fig. 3.3a). For SVM, it represents samples with small margins to the separating hyperplane (see Fig. 3.3b). Generally, a larger overlapping region of class-conditional pdfs means more potentially misclassified samples. Hence, similar classification costs of RO-TWSVM and RO-SVM can also be

Figure 3.3: Approximated class-conditional probability density functions of (a) Distance difference $d(x)$ of TWSVM and (b) SVM outputs for synthetic dataset 1.

observed from the similar size of the overlapping regions in the class-conditional pdfs, as shown in Fig. 3.3a and Fig. 3.3b, respectively.

3.3.2 Synthetic data 2: “Cross Planes” data

To show that RO-TWSVM has an advantage over RO-SVM [23] for a specific type of data due to the flexibility offered by the nonparallel hyperplanes of TWSVM, we apply these two methods on a synthetic dataset called “Cross Planes” which is adopted in some TWSVM related papers [74, 75]. In this dataset, two classes of samples are obtained by perturbing samples initially lying on two intersecting planes (lines) as shown in Fig. 3.4a.

Table 3.3: The expected costs for different cost models for “Cross Planes” dataset.

	C3	C4	C5	C6
RO-TWSVM	-9.3467	-6.7400	-3.0800	-2.4400
RO-SVM	-3.7867	-0.7533	3.3333	15.5800

In this simulation, a linear kernel is applied and the cost settings follow Table 3.1 in Section 4.1 (see also [23]). For C1 and C2, no reject region is generated due to the violation of condition (2.24) and two level curves intersecting at the same point, respectively. Since the experimental results are similar for cases C3 through C6, we only analyze the case for C5 by observing the reject regions generated by RO-TWSVM and RO-SVM (shown in Fig. 3.4b and Fig. 3.4c) and the corresponding expected costs recorded in Table 3.3. From Fig. 3.4b, it is noted that RO-TWSVM generates multiple reject regions which cover most of the sample regions with low classification confidence. In contrast, the reject region generated by RO-SVM in Fig. 3.4c is not reasonable as it contains many samples which should be classified correctly with high confidence. From Table 3.3, we observe that the proposed RO-TWSVM produces a lower expected cost than that of RO-SVM for each of the four cost models. Moreover, for the cases of C5 and C6 which have imbalanced costs associated with misclassification (false positive and false negative), the reject regions generated by RO-TWSVM are more robust since they

Figure 3.4: (a) "Cross Planes" dataset. (b) The reject region generated by the proposed RO-TWSVM. (c) The reject region generated by RO-SVM.

have smaller fluctuation than RO-SVM in the expected costs. In summary, the proposed RO-TWSVM can generate more reasonable and robust reject region than RO-SVM for datasets like “Cross Planes”.

On the one hand, this improved performance of RO-TWSVM can be explained by different formulations of SVM and TWSVM. RO-TWSVM trains two nonparallel hyperplanes corresponding to the two classes, respectively. It provides flexibility to fit the data and generate a reasonable reject region accordingly as shown in Fig. 3.4b. In contrast, for RO-SVM, the reject region is delimited by two hyperplanes which are parallel to the separating hyperplane [23] (see Fig. 3.4c). From Fig. 3.4c, it is observed that the two classes of data in “Cross Planes” cannot be linearly discriminated by one hyperplane from SVM. Therefore, RO-SVM generates an inappropriate reject region which is close to the separating hyperplane. In summary, we conclude that RO-TWSVM outperforms RO-SVM due to its flexibility offered by the nonparallel hyperplanes of TWSVM.

On another hand, the class-conditional pdfs for $d(\mathbf{x})$ and SVM outputs are also plotted in Fig. 3.5. Conclusions can be drawn from a comparison with the original sample location as shown in Fig. 3.4a. From Fig. 3.4a, it is observed that samples from both the classes in the small “intersection area” are difficult to identify and should be rejected. Fig. 3.5a illustrates that $d(\mathbf{x})$ applied by the proposed RO-TWSVM can reflect the samples in the “intersection area” with a small overlapping region of class-conditional pdfs. However, the class-conditional pdfs of SVM outputs cannot reflect this situation and even have a heavy overlapping region (see Fig. 3.5b). As a result, RO-TWSVM rejects samples that are mostly in the “intersection area” and RO-SVM mistakenly rejects a large number of samples which should be classified with high confidence. Hence, the analysis on class-conditional pdfs support the idea that the reject region determined by RO-TWSVM is more reasonable than that of RO-SVM.

(a)

(b)

Figure 3.5: Approximated class-conditional probability density functions of (a) Distance difference $d(\mathbf{x})$ of TWSVM and (b) SVM outputs for “Cross Planes” dataset.

3.4 Experiments on real-world benchmark datasets

To show the effectiveness of the proposed RO-TWSVM on real-world datasets, two groups of experiments are conducted to evaluate the effect of incorporating a reject option in TWSVM in a statistical sense. Following [23], the Wilcoxon rank sum tests are applied on 4 real-world datasets to compare the proposed RO-TWSVM with TWSVM without a reject option (TWSVM-without-RO) and the existing RO-SVM [23]. Subsequently, the Friedman and Nemenyi post-hoc tests are conducted on 23 standard binary classification datasets to statistically compare the performance of the proposed RO-TWSVM and competing methods in view of classification costs.

3.4.1 Wilcoxon rank sum test

Table 3.4: Summary of experimental datasets for Wilcoxon rank sum test.

Name	Number of Samples	%Positive	%Negative
Pima	768	34.90%	65.10%
German Credit	1000	30.00%	70.00%
Breast Cancer Wisconsin	683	34.99%	65.01%
Heart Disease Cleveland	297	27.95%	72.05%

Table 3.5: The cost models for the experiments on real-world datasets.

	CTP	CFP	CTN	CFN	CR
CM1	Unif [-10,0]	Unif [0,50]	Unif [-10,0]	Unif [0,50]	1
CM2	Unif [-10,0]	Unif [0,100]	Unif [-10,0]	Unif [0,50]	1
CM3	Unif [-10,0]	Unif [0,50]	Unif [-10,0]	Unif [0,100]	1
CM4	Unif [-10,0]	Unif [0,50]	Unif [-10,0]	Unif [0,50]	Unif [0,30]

The notation Unif $[a, b]$ denotes the uniform distribution over the interval $[a, b]$.

In this subsection, we conduct the Wilcoxon rank sum test on 4 real-world datasets [23] listed in Table 3.4. The Wilcoxon rank sum test is a non-parametric method for comparison of two classifiers over multiple datasets. In the Wilcoxon rank sum test, the pair of classifiers of interest are compared through counting the number of runs (out of

1000 runs) of wins, losses and ties based on their expected cost results. Following the settings in [23], the cost models (CM) are shown in Table 3.5. It is discussed in [23] that these 4 cost models are adopted to reduce the bias of experimental comparisons with respect to certain particular cost values. Such reduction is achieved by assuming the costs being generated by different statistical distributions rather than fixed values. In our experiment, we first compare the performance between RO-TWSVM and TWSVM-without-RO. Secondly, RO-TWSVM and RO-SVM [23] are compared. The linear and RBF kernels are adopted for both TWSVM and SVM as linear and nonlinear classifiers, respectively. The SVM is implemented using the LIBSVM software package from the public domain [76].

Algorithm 2 lists our experimental steps. The results for each CM are obtained from an experiment of 1000 runs. For each run, 5 repetitions of the experiment are implemented. In each of the 5 repetitions, the entire dataset is split into three subsets: a training set, a validation set and a testing set. The training and validation sets which account for 75% of the total samples are applied in Procedure 1 to determine the optimal thresholds t_1 and t_2 for RO-TWSVM. The remaining 25% of the samples are for testing where the expected costs are calculated. Hence, for each run, we obtain 5 expected costs over the testing set. The final cost is obtained as the mean value of these 5 expected costs. RO-SVM is run in a similar fashion.

Table 3.6 through Table 3.9 record the experimental results for all cost model and dataset pairs for linear and RBF kernel, respectively. Each cell of these tables contains three values indicating the number of runs that RO-TWSVM produces higher, lower or indistinguishable (or tie) cost compared with that of TWSVM-without-RO or RO-SVM. The indistinguishable cost is set for the case that two compared methods produce the same cost or any one of them is recorded as “RO is not applicable” [23]. For example, in Table 3.6, the cell in the upper left corner of the table indicates that for the Pima dataset tested on CM1, we obtain 19 runs that RO-TWSVM produces higher cost than TWSVM-without-RO, 865 runs that RO-TWSVM produces lower cost than TWSVM-without-RO

Algorithm 2: The experimental procedure for Wilcoxon rank sum test.

Input:
Data: $\mathbf{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$;
Output:
 $[Num_H, Num_L, Num_{ID}]$;

- 1 **Initialization:**
- 2 $Num_H \leftarrow 0$;
- 3 $Num_L \leftarrow 0$;
- 4 $Num_{ID} \leftarrow 0$;
- 5 **for** $m = 1 : 1000$ **do**
- 6 Cost Generation: CTN, CFP, CFN, CTP, CR ;
- 7 $TotalCost \leftarrow 0$;
- 8 $AverageCost \leftarrow 0$;
- 9 **for** $l = 1 : 5$ **do**
- 10 Randomly divide $\mathbf{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ into a set P_k for training and validation in Procedure 1 and a testing set P_k , respectively;
- 11 Output $\leftarrow$ Run Procedure 1 with input $(\mathbf{x}_j, y_j)$, $j \in P_k$, CTN, CFP, CFN, CTP, CR ;
- 12 **if** Output == Message: “RO is not applicable” **then**
- 13 $Num_{ID} \leftarrow Num_{ID} + 1$;
- 14 **break**;
- 15 **else if** Output == reject region $[t_1, t_2]$ **then**
- 16 Calculate TPR, TNR, FPR, FNR, RP and RN over testing samples (x_j, y_j) , $j \in P_k$;
- 17 Calculate the costs C ;
- 18 $TotalCost \leftarrow TotalCost + C$;
- 19 $AverageCost = \frac{TotalCost}{5}$;
- 20 **if** RO-TWSVM produces a higher $AverageCost$ than that of the competing method **then**
- 21 $Num_H \leftarrow Num_H + 1$;
- 22 **else if** RO-TWSVM produces a lower $AverageCost$ than that of the competing method **then**
- 23 $Num_L \leftarrow Num_L + 1$;
- 24 **else if** RO-TWSVM produces an equal $AverageCost$ to that of the competing method **then**
- 25 $Num_{ID} \leftarrow Num_{ID} + 1$;
- 26 **return** $[Num_H, Num_L, Num_{ID}]$;

(i) Num_H, Num_L, Num_{ID} denote the numbers of runs that RO-TWSVM produces higher, lower or indistinguishable (or tie) costs, respectively.
(ii) For the comparison between RO-TWSVM and RO-SVM, if any of them outputs “RO is not applicable”, we just record this case as the indistinguishable costs.

and 116 runs that the costs are indistinguishable. Therefore, we can conclude that on the basis of the expected costs, among all 1000 runs, RO-TWSVM outperforms TWSVM-without-RO for 856 runs, TWSVM-without-RO outperforms RO-TWSVM for 19 runs

and the costs are indistinguishable for the rest of 116 runs.

Table 3.6: Comparison between RO-TWSVM and TWSVM-without-RO for linear kernel based on Wilcoxon rank sum test.

		Pima	German Credit	Breast Cancer	Cleveland
CM1	Higher Cost	19	81	4	35
	Lower Cost	865	798	821	844
	Indistinguishable	116	121	175	121
CM2	Higher Cost	13	45	4	12
	Lower Cost	909	860	847	903
	Indistinguishable	78	95	149	85
CM3	Higher Cost	7	35	2	16
	Lower Cost	910	894	878	900
	Indistinguishable	83	71	120	84
CM4	Higher Cost	45	124	2	45
	Lower Cost	241	125	224	207
	Indistinguishable	714	751	774	748

(i) Each cell of the table contains three values which indicate the number of runs that RO-TWSVM produces higher, lower or indistinguishable (or tie) cost compared with that of TWSVM-without-RO or RO-SVM. The second value highlighted in each cell is the number of runs that RO-TWSVM outperforms its competing method. This presentation format follows that used in [23].

Table 3.7: Comparison between RO-TWSVM and TWSVM-without-RO for RBF kernel based on Wilcoxon rank sum test.

		Pima	German Credit	Breast Cancer	Cleveland
CM1	Higher Cost	51	87	3	267
	Lower Cost	833	792	797	590
	Indistinguishable	116	121	200	143
CM2	Higher Cost	23	50	5	379
	Lower Cost	900	857	814	504
	Indistinguishable	77	93	181	117
CM3	Higher Cost	23	39	3	128
	Lower Cost	894	890	862	765
	Indistinguishable	83	71	135	107
CM4	Higher Cost	70	109	6	115
	Lower Cost	217	139	207	141
	Indistinguishable	713	752	787	744

Table 3.6 and Table 3.7 record the comparison results between RO-TWSVM and

TWSVM-without-RO. From these tables, we observe the number of runs that RO-TWSVM produces a lower classification cost is significantly more than the numbers of the other two outcomes. Specifically, for the cost models CM1 through CM3, the number of runs that RO-TWSVM outperforms TWSVM-without-RO is clearly dominant. For CM4, the cost of rejection follows the distribution $Unif[0,30]$ rather than the fixed value. Therefore it is more probable that condition (2.24) is not satisfied which leads to the increasing percentage where the comparison result “RO is not applicable” occurs. In summary, RO-TWSVM outperforms TWSVM-without-RO in view of reduction of classification cost.

Table 3.8: Comparison between RO-TWSVM and RO-SVM for linear kernel based on Wilcoxon rank sum test.

		Pima	German Credit	Breast Cancer	Cleveland
CM1	Higher Cost	326	395	642	302
	Lower Cost	564	488	266	589
	Indistinguishable	110	117	92	109
CM2	Higher Cost	213	356	602	355
	Lower Cost	712	559	331	575
	Indistinguishable	75	85	67	70
CM3	Higher Cost	285	413	732	260
	Lower Cost	638	519	198	665
	Indistinguishable	77	68	70	75
CM4	Higher Cost	156	129	223	113
	Lower Cost	135	125	59	159
	Indistinguishable	709	746	718	728

Table 3.8 and Table 3.9 record the comparison results between RO-TWSVM and RO-SVM. From these tables we observe for the datasets German Credit and Cleveland, our proposed RO-TWSVM outperforms RO-SVM for C1 through C4 and for both kernels. For the Pima dataset, RO-TWSVM outperforms RO-SVM dominantly using a linear kernel and achieves comparable performance using an RBF kernel. For the Breast Cancer dataset, RO-SVM produces a lower classification cost for all the cost models for both kernels. However, we notice that for the Breast Cancer dataset,

Table 3.9: Comparison between RO-TWSVM and RO-SVM for RBF kernel based on Wilcoxon rank sum test.

		Pima	German Credit	Breast Cancer	Cleveland
CM1	Higher Cost	458	127	507	230
	Lower Cost	432	756	401	661
	Indistinguishable	110	117	92	109
CM2	Higher Cost	506	155	424	292
	Lower Cost	419	760	509	638
	Indistinguishable	75	85	67	70
CM3	Higher Cost	469	98	667	118
	Lower Cost	454	834	263	807
	Indistinguishable	77	68	70	75
CM4	Higher Cost	150	53	126	108
	Lower Cost	141	201	156	164
	Indistinguishable	709	746	718	728

TWSVM and SVM both achieve a high classification rate (above 95%) which shows this dataset is quite separable. In fact, for a separable dataset, the reject option setting may not be necessary because standard classifiers without RO could already achieve high classification accuracy. In summary, for datasets which are not quite separable, based on the Wilcoxon rank sum test, our proposed RO-TWSVM outperforms RO-SVM in two out of four datasets (German and Cleveland) while the two methods are comparable for one dataset (Pima). However, for the dataset (Breast Cancer) which is relatively separable by one hyperplane, RO-SVM outperforms the proposed RO-TWSVM.

3.4.2 Friedman and Nemenyi post-hoc test

In this subsection, we compare the proposed RO-TWSVM with 3 competing methods, namely TWSVM-without-RO, RO-SVM [23] and SVM-without-RO (i.e., SVM without a reject option), on 23 standard binary classification datasets listed in Table 3.10 [77–79]. The adopted cost models follow those listed in Table 3.5. For each dataset and cost model (CM1 through CM4), 10 expected costs for each method are calculated using a linear and an RBF kernel, respectively. To get a clear picture of the performance with

respect to each dataset, the average values of the expected costs are plotted in Fig. 3.6a and Fig. 3.6b for classifiers with linear and RBF kernel, respectively. It is observed that, on average, RO-TWSVM can produce a lower expected cost than both TWSVM-without-RO and SVM-without-RO. This observation further verifies the conclusion we obtain in Section 5.1 that embedding a reject region can enhance the robustness of the corresponding classifiers. Comparison results between RO-TWSVM and RO-SVM are observed to be dependent on the adopted kernels. When the linear kernel is adopted, RO-TWSVM and RO-SVM achieve a comparable performance in view of average costs (see Fig. 3.6a). In contrary, RO-TWSVM can produce a lower classification cost than RO-SVM when the RBF kernel is adopted (see Fig. 3.6b). From these figures, we conclude that, on average, RO-TWSVM outperforms RO-SVM based on the reduction of the classification cost.

Since the average cost is a general criterion to measure the performance of the methods, we conduct a Friedman test together with a Nemenyi test to verify the statistical significance of experimental conclusions. The Friedman test is a non-parametric method to compare multiple methods over multiple datasets based on their average ranks [80]. The null hypothesis of Friedman test is that all 4 methods are equivalent based on their average ranks. In our experiment, the method with the lowest classification cost is ranked as 1, the second lowest is ranked as 2, and so on. If the methods produce the same cost, an average rank is assigned to each of them. The confidence level of the Friedman test is set as 95% where the null hypothesis can be rejected if the calculated p value less than 0.05. A post-hoc analysis called Nemenyi test [80] is subsequently conducted to evaluate whether the differences in the average ranks are statistically significant among the compared methods. For the Nemenyi test, the expected costs of any two methods are significantly different if the corresponding average ranks differ by at least the Critical Difference (CD) [80]. Otherwise, the two methods have no significant difference in performance based on their classification costs. The graphical representations of Nemenyi test results are shown in Fig. 3.7 and Fig. 3.8, where the methods that are

Table 3.10: Summary of experimental datasets for Friedman and Nemenyi tests.

	Name	Number of Samples	%Positive	%Negative
1	Australian Credit	690	44.49%	55.51%
2	Breast Cancer Wisconsin	683	34.99%	65.01%
3	Cleveland	297	27.95%	72.05%
4	CMC	1437	57.30%	42.70%
5	German Credit	1000	30%	70%
6	Glass	214	23.83%	76.71%
7	Haberman	306	26.47%	73.53%
8	Heart Statlog	270	44.44%	53.87%
9	Ionosphere	351	64.01%	35.09%
10	Liver	345	57.97%	42.03%
11	Monk1	124	50%	50%
12	Monk2	169	37.87%	62.13%
13	Monk3	122	49.18%	50.82%
14	Mushroom	5644	61.80%	38.20%
15	Pima	768	34.90%	65.10%
16	shuttle	279	52.33%	47.67%
17	Sonar	208	53.37%	46.63%
18	Splice	1000	51.70%	48.30%
19	Tic-tac-toe	958	34.66%	65.34%
20	Vote	435	61.38%	38.62%
21	wdbc	569	62.74%	37.26%
22	Wpbc	194	23.71%	76.29%
23	Yeast1	1484	28.91%	71.09%

not significantly different in performance are connected. In the following, we show the results of two scenarios which consist of all methods adopting a linear kernel and an RBF kernel, respectively. In these figures, the numbers in the brackets are the specific average ranks of the corresponding methods.

From Fig. 3.7 and Fig. 3.8, it is first observed that both RO-TWSVM and RO-SVM have significantly higher average ranks (i.e., with lower rank value) than TWSVM-without-RO and SVM-without-RO, respectively. These results demonstrate that incorporating a reject option can significantly improve the performance of the original classifiers in terms of reducing classification costs. Secondly, the Nemenyi test results between RO-TWSVM and RO-SVM are observed to be different for different kernels. Using the linear kernel, there is no significant difference between the classification costs of

(a)

(b)

Figure 3.6: The average cost for each dataset : (a) linear kernel , (b) RBF kernel.

RO-TWSVM and RO-SVM. In contrast, in the case of using RBF kernel, RO-TWSVM shows statistically better performance than RO-SVM by means of the Nemenyi test. Therefore, we can conclude that, for these datasets, RO-TWSVM achieves a comparable performance with RO-SVM when using the linear kernel. Whereas using the RBF kernel, RO-TWSVM significantly outperforms RO-SVM in view of classification costs.

Remark. According to [80], the Wilcoxon and Friedman test are non-parametric methods suitable for statistical evaluation of classifiers. These two tests are particularly

(a)

(b)

(c)

(d)

Figure 3.7: Comparison results of four methods using linear kernel through Nemenyi test: (a) CM1, (b) CM2, (c) CM3 and (d) CM4.

Figure 3.8: Comparison results of 4 methods using RBF kernel through Nemenyi test: (a) CM1, (b) CM2, (c) CM3 and (d) CM4.

Table 3.11: Summary of the experimental results

Synthetic data sets
<ol style="list-style-type: none"> 1. The proposed RO-TWSVM can produce reject regions with different sizes and locations according to various cost settings. 2. For synthetic data set 1 in Section 3.3.1, RO-TWSVM and RO-SVM generate similar reject regions and achieve similar performance in terms of classification costs. 3. For synthetic data set 2 called “Cross Planes” in Section 3.3.2, RO-TWSVM generates more reasonable and robust reject regions than RO-SVM due to the flexibility offered by the nonparallel hyperplanes of TWSVM.
Real-world data sets
<ol style="list-style-type: none"> 1. The proposed RO-TWSVM significantly outperforms TWSVM-without-RO in terms of classification costs for all the datasets based on the statistical tests in both Section 3.4. 2. In Section 3.4.1, through Wilcoxon rank sum test on 4 datasets, the performance comparison between RO-TWSVM and RO-SVM is dataset dependent. For the datasets which can be easily discriminated by one hyperplane, the two methods achieve comparable performance. Whereas, RO-TWSVM significantly outperforms RO-SVM for data sets which are difficult to be discriminated by one hyperplane. 3. In Section 3.4.2, four methods, namely, the proposed RO-TWSVM, the TWSVM-without-RO, the RO-SVM and the SVM-without-RO, are compared on 23 real-world datasets. The comparison is based on average costs for each dataset and Friedman test together with Nemenyi test. As for the comparison between RO-TWSVM and RO-SVM, while the two methods achieve comparable performance using a linear kernel, RO-TWSVM significantly outperforms RO-SVM when using an RBF kernel.

applicable for our experiments since the expected costs for different datasets are comparable (in another word, commensurability [80]). Also, they do not assume normal distributions or homogeneity of variance for the compared random variables (in our case, the expected costs) [80]. These relaxed assumptions ensure the generality of the tests. Hence, it is appropriate that we conduct the Wilcoxon rank sum test instead of the paired t-test [5] when comparing two methods. As for comparing multiple methods over multiple datasets, the Friedman and Nemenyi tests are recommended [80]. In our experiments, we attempt to include all the available binary classification datasets from the related papers [77–79] to reduce as much as possible the impact of dataset dependency.

3.5 Summary

In this chapter, we have proposed a twin support vector machine with a reject option (RO-TWSVM) through the ROC curve. The method is adaptable to changes in cost settings which is a desirable property in real-world applications. Similar to RO-SVM [23], RO-TWSVM is also formulated based on a cost-sensitive framework which follows the principle of minimization of the expected cost of classification. Hence, on the one hand, some advantages of RO-SVM are retained in the proposed RO-TWSVM, such as the training for TWSVM hyperplanes and the plot of ROC curve are implemented only once since changes made on cost settings, if any, only influence the slopes of the level curves. Hence, excessive re-calculations for the TWSVM hyperplanes and the ROC curve are not required. On the other hand, RO-TWSVM does not assume that samples being rejected should stay close to the separating hyperplane as RO-SVM does. Since we choose the distance difference in TWSVM to measure the uncertainty of a sample, an outlier sample which is far away from the SVM hyperplane and not rejected by RO-SVM could still be rejected by the proposed RO-TWSVM. The experimental results on two synthetic datasets and multiple real-world datasets have shown that RO-TWSVM

significantly outperforms TWSVM-without-RO, and in general, performs better than RO-SVM for datasets which are difficult to be discriminated by one hyperplane. Table 3.11 summarizes the major experimental results for both synthetic and real-world data sets.

Different data set has different distribution patterns for the ambiguous and outlier samples. In general, ambiguous samples stay close to the decision boundary and they could be selected by our RO-TWSVM if the cost values are properly designed since the distance difference of an ambiguous sample to both TWSVM hyperplanes is small. Outlier samples have more possible distributions. For those outlier samples which are far away from both hyperplanes generated by RO-TWSVM, they could also be selected by RO-TWSVM since they also have the similar distances to both hyperplanes. However, our RO-TWSVM cannot select all the outlier samples since the hyperplanes to represent the training samples might not be sufficiently reliable in the feature space with small number of training samples. Hence, the RO-TWSVM can better select both ambiguous and outlier samples if more training samples are available.

Chapter 4

A two-stage method for robust biomedical image classification

Chapter 4 proposes a two-stage method for robust biomedical image classification. Section 4.1 summarizes major challenges of biomedical image classification as the motivation of the proposed method. In Section 4.2, we present the technical details of the proposed method. Section 4.3 presents and analyzes experimental results on three real-world benchmark biomedical image datasets. Section 4.4 summarizes this chapter.

4.1 Introduction

As discussed in Section 2.4.1, biomedical image classification could confront the challenges of high level noise, image blur, illumination variation and complicated geometric correspondence among various categorical biomedical patterns in practice. These challenges are difficult to be handled simultaneously by using a single classification model trained on images represented based on a single local feature.

To alleviate the aforementioned challenges, we propose a two-stage method for robust biomedical image classification. At stage 1, we propose a confidence score based classification rule with a reject option for a preliminary decision using the support vector machine (SVM). The testing images going through stage 1 are separated into two groups based on their confidence scores. Those testing images with sufficiently high confidence scores are classified at stage 1 while the others with low confidence scores are rejected and fed to stage 2. At stage 2, the rejected images from stage 1 are first processed by a subspace analysis technique called eigenfeature regularization and extraction (ERE), and then classified by another SVM trained in the transformed subspace learned by ERE. At both stages, images are represented based on two types of local features, i.e., SIFT and SURF, respectively. They are encoded using various bag-of-words (BoW) models to handle biomedical patterns with and without geometric correspondence, respectively.

The robustness of the proposed method is achieved by the complementary effects from three perspectives. Firstly, the classification structure is complementary via a reject option method, i.e., the methods of both stages are formulated to classify images with high confidence. Secondly, the local features adopted, namely SIFT and SURF, have complementary merits in their robustness to variations in images. Thirdly, the image representation via different BoW models regarding whether incorporating spatial information is also complementary. In general, the proposed two-stage structure fuses the two complementary image representations via a reject option method and an ERE subspace analysis technique. Thereby, the proposed method is developed to classify biomedical patterns with high confidence.

Extensive experiments are conducted on three real-world benchmark datasets of biomedical images. First, The complementary effects discussed above are studied by comparing the proposed method with the individual method of each stage. It is shown that the proposed method makes the methods of the two stages flexible to simultaneously handle the images with different geometric correspondence. Second, by comparing with several state-of-the-art related methods, it is shown that the proposed method outperforms the competing methods in terms of multiple classification performance metrics.

4.2 Methodology

4.2.1 Overview of the proposed method

As discussed in Section 4.1, the proposed method consists of two stages which are complementary in their classification goals. At stage 1, we propose a support vector machine with a reject option trained on the images represented by high-dimensional SIFT feature based bag-of-words (BoW) models. The model incorporates spatial information using spatial pyramid matching (SPM) [49] reviewed in Section 2.3.1. Motivated by the Chow's rule of classification with a reject option [4] reviewed in Section 2.1.1, we propose a confidence score (CS) based classification rule with a reject option for the SVM. For a testing image being classified with a high score, it is assigned to its identified category at stage 1. Otherwise, the image would be rejected and fed into stage 2 for further analysis. At stage 2, the training images are represented by SURF feature based BoW models without incorporating spatial information. A subspace analysis technique called eigenfeature regularization and extraction (ERE) [58] (reviewed in Section 2.3.2) is applied on the BoW models to learn a discriminative subspace. After that, another SVM is trained in the transformed subspace learned by ERE. The training phases of the two stages are implemented first over the same training images. In the testing phase, a testing image will be either classified at stage 1 with high

confidence or rejected and classified at stage 2.

Fig. 4.1 shows the flowchart of classifying a batch of testing images using the proposed two-stage method. The testing images go through the two stages in sequence. At stage 1, for those testing images classified by the first SVM with high confidence scores, their identified categories are assigned. However, for those testing images which are rejected due to low confidence scores, they are fed into stage 2 and being classified in the transformed subspace by the second SVM. The following subsections present the details of the methods at the two stages, respectively.

Figure 4.1: The classification flowchart of the proposed method for a batch of testing images.

4.2.2 Stage 1: Support vector machine with a reject option

The goal of stage 1 is to train an SVM with a reject option. The testing images going through this stage are separated into two groups. Those testing images with sufficiently high classification confidence are classified at stage 1 while the others with low confidence are rejected and fed into stage 2. To decide whether a testing image should be classified or rejected, we propose, in both the training phase and testing phase at stage 1, a confidence score (CS) based classification rule with a reject option for SVM. The confidence scores are defined for a testing sample based on its distances to the SVM decision hyperplanes and to the sample associated categorical centroids. This definition is an extension from our preliminary work [81] on similarity scoring for microscopy image classification.

Distance based confidence score of SVM classification

In this subsection, a distance based confidence score (CS) for SVM is formulated in the context of an SVM with a linear kernel. The formulation can be extended to a nonlinear kernel SVM by using a nonlinear mapping function $\phi(\mathbf{x})$ instead of the original feature vector $\mathbf{x}$ adopted here.

In binary classification, SVM solves for an optimal decision hyperplane to discriminate samples into two classes. The hyperplane can be parameterized as:

$$\mathbf{w}^T \mathbf{x} + b = 0, \quad (4.1)$$

where $\mathbf{w}$ is a normal vector to the hyperplane and b is a bias term. The optimal hyperplane is generally obtained by solving a quadratic programming problem:

$$\begin{aligned} & \underset{\mathbf{w}, b}{\text{minimize}} && \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi_n, \\ & \text{subject to} && y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n, \\ & && \xi_n \geq 0, \end{aligned} \quad (4.2)$$

where $\mathbf{x}_n$ and $y_n \in \{-1, +1\}$ ($n = 1, 2, \dots, N$) represent the sample vector and the categorical label of the n_{th} training image, respectively. ξ_n is a slack variable introduced for each training sample and a regularization parameter C controls the trade-off between the sum of all the slack variables and the margin width. $\|\cdot\|$ represents the Euclidean norm. For multi-categorical classification, we apply the one-versus-rest strategy for SVM training, i.e., training one decision hyperplane per class, with the training samples of that class as the positive samples and the remaining samples as the negatives.

Once the decision hyperplanes are obtained, a distance based confidence score proposed in [81] is adopted to quantify the confidence level of a testing sample for each category. The confidence score (CS) is calculated as the average value between a discriminative score (DS) and a prior score (PS). In particular, for a testing sample, DS shows the confidence level based on its distances to the hyperplanes while PS shows

the confidence level in view of its distances to the categorical centroids. Several distance notations are illustrated in Fig. 4.2. The calculation steps for CS associated with a testing sample $\mathbf{x}_t$ are shown as follows.

Figure 4.2: A diagram illustrating an example which contains 3 categories of samples shown by the blue triangles, red circles and purple squares, respectively. In this figure, we only plot the hyperplane to discriminate the samples of the category shown by the blue triangles. The signed distance D_h is defined as the perpendicular distance between a testing sample (denoted by the green diamond) and the SVM hyperplane with the sign determined by the SVM output. D_h is adopted in the calculation of the discriminative score for blue triangle category. The discriminative scores for the other category can be calculated in a similar way. D_c^m ($m = 1, 2, 3$) denotes the distance between the testing sample and the m_{th} category centroid and they are applied to calculate the prior scores.

1. Calculate $DS_m(\mathbf{x}_t)$ ($m = 1, 2, \dots, M$) associated with the m_{th} category. Define the signed distance $D_h^m(\mathbf{x}_t) = \frac{\mathbf{w}_m^T \mathbf{x}_t + b_m}{\|\mathbf{w}_m\|}$, where $\mathbf{w}_m$ and b_m are the parameters associated with the decision hyperplane for the m_{th} category. Define $DS_m(\mathbf{x}_t) = \frac{1}{1 + e^{-\alpha D_h^m(\mathbf{x}_t)}}$, where the value of α is set at 10. The sigmoid function is adopted to constrain the range of $DS_m(\mathbf{x}_t)$ within $[0, 1]$.
2. Calculate $PS_m(\mathbf{x}_t)$ ($m = 1, 2, \dots, M$) associated with the m_{th} category. Define the distance between $\mathbf{x}_t$ and the m_{th} categorical centroid as $D_c^m(\mathbf{x}_t)$ ($m = 1, 2, \dots, M$). Calculate $PS_m(\mathbf{x}_t) = \frac{\sum_{n=1, n \neq m}^M D_c^n(\mathbf{x}_t)}{\sum_{m=1}^M D_c^m(\mathbf{x}_t)}$.
3. Normalize the above DS and PS with respect to the categories, i.e., calculate

$$DS_m(\mathbf{x}_t)^{\text{norm}} = \frac{DS_m(\mathbf{x}_t)}{\sum_{m=1}^M DS_m(\mathbf{x}_t)} \text{ and } PS_m(\mathbf{x}_t)^{\text{norm}} = \frac{PS_m(\mathbf{x}_t)}{\sum_{m=1}^M PS_m(\mathbf{x}_t)}.$$

4. Calculate the confidence score $CS_m(\mathbf{x}_t)$ ($m = 1, 2, \dots, M$) associated with $\mathbf{x}_t$ and the m_{th} category by $CS_m(\mathbf{x}_t) = \frac{(DS_m(\mathbf{x}_t)^{\text{norm}} + PS_m(\mathbf{x}_t)^{\text{norm}})}{2}$.

The confidence score (CS) is a reasonable index to quantify the confidence of the classification of a testing sample with respect to each category. In particular, a high confidence score reflects either the sample stays far away from the decision hyperplanes indicated by DS , or close to one of the class centroids indicated by PS . On the contrary, the testing samples with low confidence scores are mostly near the decision hyperplanes or far away from the class centroids. In our method, the samples with low confidence scores are to be rejected via a reject option at stage 1. The formulation of a CS based classification rule with a reject option is proposed in the next subsection.

The confidence score based classification rule with a reject option

The method at Stage 1 is based on modifying Chow's rule (reviewed in Section 2.1.1) using the proposed confidence score. Assuming that $c_1, c_2, \dots, c_M$ denote the categorical labels for an M-categorical classification problem, according to the original Chow's rule [4], a sample $\mathbf{x}$ is:

- rejected whenever

$$\max_{c_i} p(c_i|\mathbf{x}) < (1 - t), \quad (4.3)$$

- or accepted and labeled to the category c_j whenever

$$\begin{aligned} p(c_j|\mathbf{x}) &\geq (1 - t), \\ p(c_j|\mathbf{x}) &> p(c_i|\mathbf{x}) \quad \forall c_i, i \neq j, \end{aligned} \quad (4.4)$$

where $p(c_i|\mathbf{x})$ is the posterior probability of the testing sample $\mathbf{x}$ and it can be estimated associated with certain classification model (see, e.g., [24]). The parameter t is a constant within $[0, 1]$. In this work, we propose to approximate the posterior probabilities $p(c_i|\mathbf{x})$

using the confidence scores $CS_i(\mathbf{x})$. Based on the analysis in our previous work [81], we concluded that the proposed confidence score can generally better evaluate the classification results of biomedical images. Hence, in this work, we apply Chow's rule through replacing the true posterior probability by our confidence score to link the two stages which contain the complementary features. Specifically, a testing sample $\mathbf{x}$ is

- rejected whenever

$$\max_i CS_i(\mathbf{x}) < (1 - t), \quad (4.5)$$

- or accepted and labeled to the category c_j whenever

$$\begin{aligned} CS_j(\mathbf{x}) &\geq (1 - t), \\ CS_j(\mathbf{x}) &> CS_i(\mathbf{x}) \quad \forall i, i \neq j. \end{aligned} \quad (4.6)$$

As reviewed in Section 2.1.1, the parameter t controls the trade-off between the error rate and the rejection rate of classification. Fig. 4.3 illustrates the curves of the error rate with rejection, the rejection rate and the error rate without rejection for the Hep-2 cell image dataset. These curves are plotted with respect to various values of the parameter t within $[0, 1]$. A trade-off between the error rate and the rejection rate can be observed. In particular, when the parameter t is within the region of $[0.5, 0.7]$, the trade-off is clear.

Chow's rule ensures the minimum error rate given a rejection rate. In our formulation, we calculate the optimal t by setting a rejection rate α_r . The setting of α_r is application dependent. For the dataset which is generally separable, a low α_r is set so that most samples can be classified correctly with high confidence. Otherwise, for the non-separable data, a high α_r is set to reduce misclassification by introducing more rejected samples. The optimal value of α_r is determined through cross validation. Specifically, we search from the interval of $[0.05, 0.07, \dots, 0.5]$ and determine the optimal value of α_r which produces the highest accuracy on the validation set. Once the optimal α_r is obtained, the value of t can be determined through the rejection rate curve as shown

Figure 4.3: The figure shows the error rate with rejection (blue dashed line), the rejection rate (red solid line) and the error rate without reject option (grey dotted line) with respect to t for the Hep-2 cell image dataset.

in Algorithm 3.

Figure 4.4: The procedure of stage 1. At this stage, the SIFT feature based BoW model is applied. In the training phase, an SVM with a reject option is trained. In the testing phase, the testing images are either classified with high confidence or rejected and fed to stage 2.

After t is determined, the classification rule with a reject option can be obtained. In particular, given a testing sample $\mathbf{x}_t$, the values of $CS_m(\mathbf{x}_t) (m = 1, 2, \dots, M)$ are calculated

Algorithm 3: The procedure of determining t **Input:**

Data: $\mathbf{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i$ and y_i denote the feature vector of the i_{th} training image and its categorical label, respectively;

The rejection rate: α_r ;

Output:

The value of t ;

- 1 Randomly divide $\mathbf{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ into a training set P_k and a validation set P_k , respectively.
- 2 Use the samples in the training set P_k to train for an SVM model.
- 3 Calculate the confidence scores for each sample in the validation set P_k .
- 4 Determine the rejection rate curve with respect to t within the range of $[0, 1]$ using the confidence scores for the samples in the validation set P_k .
- 5 Based on the rejection rate α_r , determine the corresponding t from the rejection rate curve.

and the rules of (4.5) and (4.6) are applied to decide if the sample $\mathbf{x}_i$ is classified at stage 1 or rejected and fed to stage 2. In other words, the reject option provides a criterion for separating the testing samples into two classification stages based on their confidence scores. Fig. 4.4 summarizes the procedure of stage 1.

Remark. For Stage 1, attempts have been made to exploit the RO-TWSVM (proposed in Chapter 3) to incorporate a reject option. However, through experiments, we observe that TWSVM model has difficulty to handle high dimensional SPM based BoW feature vectors. Also, for general biomedical image classification, it is hard to determine the optimal cost settings as discussed in Chapter 3. Hence, in this chapter, we propose a confidence score based support vector machine with a reject option at Stage 1. Without considering costs, we tune only one parameter, i.e., the rejection rate, to control the relationship between classification and rejection.

4.2.3 Stage 2: Subspace analysis of the rejected images

Stage 2 of the proposed method is to classify the rejected images from stage 1. At this stage, the subspace analysis method called eigenfeature regularization and extraction (ERE) [58] is applied to find a discriminative subspace to classify the rejected samples. The goal of ERE [58] at this stage is to extract the discriminative structures underlying

a small number of training samples. At stage 2, the SURF feature based BoW models are adopted. To visualize the effectiveness of ERE for our application, we compare the 2D scatter plot after ERE with that after the popular dimensionality reduction technique called principal component analysis (PCA). Figs. 4.5a and 4.5b illustrate the samples from the 6-class Hep-2 cell image dataset after feature dimensionality reduction. In Fig. 4.5a, the PCA is applied on the original feature vectors associated with the most significant 2 eigenvalues. In Fig. 4.5b, the samples are transformed by ERE into the 2-dimensional subspace. By comparing these two figures, it is clearly seen that the ERE transformed features are more discriminative in view of smaller overlapping among the categorical samples.

Figure 4.5: The 2D scatter plots for the Hep-2 cell image dataset with 6 categories of images: (a) the samples in the PCA transformed subspace. (b) the samples in the ERE transformed subspace.

By means of ERE, in the training phase, the following steps are carried out:

1. First, the eigenfeature regularization replaces the large eigenvalues of the within-class scatter matrix with designed values as discussed in Section 2.3.2.
2. Secondly, a discriminative subspace is obtained by eigen decomposition on the regularized total scatter matrix, represented by a transformation matrix $\mathbf{U}$ learned over the training samples.

3. Finally, a linear kernel SVM is trained over the training samples in the transformed subspace.

In the testing phase, given a rejected sample $\mathbf{x}_r$, it is firstly transformed by $\mathbf{U}$, i.e., $\mathbf{f}_r = \mathbf{U}^T \mathbf{x}_r$ and then classified by the SVM. Fig. 4.6 summarizes the procedure of stage 2.

Figure 4.6: The procedure of stage 2. At this stage, the SURF feature based BoW model is applied. In the training phase, ERE is applied on the training samples and an SVM is trained in the transformed subspace. In the testing phase, the rejected samples (images) from stage 1 are transformed to the ERE learned subspace and classified by the SVM.

To summarize, the proposed method combines the two stages described in Figs. 4.4 and 4.6, respectively, into a cascaded classification structure (see also Fig. 4.1). In the following, the extensive experiments are conducted on real-world biomedical image datasets to evaluate the performance of the proposed method compared with some state-of-the-art competing methods.

4.3 Experiments

4.3.1 Datasets

Three public benchmark biomedical image datasets, namely the 2D Hela dataset [13], the PAP smear dataset [59] and the Hep-2 cell image dataset [60] are adopted for the experiments. Since both SIFT and SURF features are extracted from grayscale images,

for those color images in the PAP smear and the Hep-2 cell image datasets, they are transformed by a weighted sum of R(red), G(green), and B(blue) channel values [43], i.e., $Grayscale = 0.2990 \times R + 0.5870 \times G + 0.1140 \times B$.

2D Hela dataset

The 2D Hela dataset [13] contains fluorescence microscope images of human Hela cells using 10 different labels to determine the distribution of different proteins within cells. It is one of the most adopted datasets for evaluating machine learning approaches to microscope image classification. It contains 10 categories of 862 fluorescence microscopic images. The details of the dataset and some sample images for each category are shown in Table 4.1 and Fig. 4.7, respectively.

Table 4.1: A summary of the 2D Hela dataset

Category	#Image per category
ActinFilaments	98
Nucleus	87
Endosomes	91
ER	86
Golgi Giantin	87
Golgi GPP130	85
Lysosome	84
Microtubules	91
Mitochondria	73
Nucleolus	80

Figure 4.7: Some sample images in the 2D Hela dataset for each category.

Table 4.2: A summary of the PAP smear dataset

Category	#Image per category
Normal	242
Abnormal	675

Figure 4.8: Some sample images in the PAP smear dataset for each category.

PAP smear dataset

The PAP smear dataset [59] contains the images collected at the Herley University Hospital by means of a digital camera and microscopes. The images are the samples of human cells stained by Papanicolau method which are for cell examination under a microscope for pre-cancer diagnosis. The dataset contains 7 different categories labeled by skilled cyto-technicians and doctors. These images are classified as normal and abnormal classes from diagnosis results. In this chapter, we are aiming at the binary classification, i.e., classifying an image as normal or abnormal. Table 4.2 shows the number of images per class and Fig. 4.8 shows the sample images of both classes.

Hep-2 cell image dataset

The Hep-2 cell image dataset [60] contains indirect immunofluorescence (IIF) images which are applied to detect autoimmune diseases by searching for antibodies in the patient serum. This dataset was used by the contest of the International Conference

on Pattern Recognition (ICPR). It has 6 categories of images cropped from 28 big IIF images. Table 4.3 and Fig. 4.9 show the numbers of images per category and some sample images, respectively.

Table 4.3: A summary of the Hep-2 cell image dataset

Category	#Image per category
Centromere	357
Homogeneous	330
Nucleolar	241
Coarse speckled	210
Fine speckled	208
Cytoplasmatic	109

Figure 4.9: Some sample images in the Hep-2 cell dataset for each category.

4.3.2 Experimental Settings

Two separate experiments are adopted on the above three real-world biomedical image datasets. The first experiment is to evaluate the effectiveness of the proposed two-stage classification structure. This is implemented by a comparison between the proposed method and the individual method at each stage, respectively. The second experiment is to compare the proposed method with those state-of-the-art methods reported in [43]. Specifically, we compare our method with the SAHLBP based method proposed in [43] and also with several SVM based methods under the bag-of-words framework using different local features. To evaluate the performance of different classifiers, we also

incorporate the classification results of two random forest [82] based methods using our designed bag-of-words features.

We conduct 30 rounds of experiments by randomly splitting the data. In each round, for each category, 80% of the total images are randomly selected as training images and the remaining 20% are used for testing. The average classification accuracies and standard deviations on the testing images from the 30 rounds of experiments are calculated. The classification accuracy is calculated by:

$$\text{Accuracy} = \frac{\# \text{ Correctly classified testing images}}{\# \text{ Total testing images}} \times 100\% \quad (4.7)$$

4.3.3 Experiment 1

In experiment 1, the proposed two-stage method is compared with each of those individual methods of both stages, respectively. To make a fair comparison, the parameters are optimally selected and set at the same values for the compared methods. In particular, the patch size of both SIFT and SURF is set at 16 pixels and the grid space of the neighboring patches is set at 4 pixels. The number of visual words in the BoW dictionary is 1000. The level of spatial pyramid matching (SPM) (applied only for stage 1) is set at 3. As the methods at both stages train linear SVMs, the parameter C is selected through a 5-fold cross validation from the range of $[2^{-15}, 2^{-14}, \dots, 2^{15}]$. In order to preserve the imbalanced prior probabilities across the training process and the testing process, we perform an evaluation based on the stratified cross validation where the distribution of images of each category in the training set and the testing set is consistent with that in the original entire dataset [43]. For the highly imbalanced dataset PAP smear, we implement an informed undersampling technique called EasyEnsemble where the majority abnormal samples are undersampled into three independent subsets and an SVM is trained on each subset of abnormal samples together with the minority

normal samples [83]. The final SVM hyperplane is calculated as the average of the three SVM hyperplanes. At stage 1 of the proposed method, we randomly select 50% of the training images for SVM training and the remaining 50% of images as a validation set to formulate a reject option. The rejection rate α_r is searched from the interval $[0.05, 0.07, \dots, 0.5]$ and the corresponding parameter t is determined by the procedure described in Algorithm 3.

Table 4.4: Classification accuracy comparison of the proposed two-stage method and the individual method of each stage.

Methods	2D Hela	PAP smear	Hep-2
Method of stage 1	89.45% $\pm$ 1.8%	88.88% $\pm$ 1.5%	91.73% $\pm$ 1.0%
Method of stage 2	91.60% $\pm$ 1.5%	88.21% $\pm$ 1.5%	91.72% $\pm$ 1.2%
The proposed method	92.96% $\pm$ 1.3%	90.96% $\pm$ 0.5%	92.97% $\pm$ 1.0%

Figs. 4.10a to 4.10c show the classification accuracies among the proposed method and the individual methods for the 30 rounds of the experiments. Quantitatively, Table 4.4 presents the average classification accuracies over the experiments. From the figures and the table, it is noted that, almost for every round, the two individual methods adopted at one separate stage produce comparable classification accuracies. However, the proposed method achieves a consistently better performance than both the individual methods. These observations show that the two individual methods at two stages are complementary and the proposed two-stage method effectively combines the two stages to improve the classification accuracy.

To have a clearer picture of the performance gain through cascading the two stages, we plot the receiver operating characteristic (ROC) curve and calculate the average values of area under the curve (AUC), the true positive rates (TPR) and the false positive rates (FPR) over the 30 rounds of experiments. These statistics are calculated for the PAP smear dataset since it is a binary classification problem where the normal images are treated as the positive class and the abnormal images are the negative class. The TPR

(a) The 2D Hela dataset

(b) The PAP smear dataset

(c) The Hep-2 cell dataset

Figure 4.10: The classification accuracies from 30 rounds of experiments for each dataset : (a) 2D Hela, (b) PAP and (3) Hep-2. The blue dashed lines denote the accuracies of the method of stage 1 and the red dotted lines show that of the method of stage 2. The green solid lines denote the accuracies of the proposed method.

and FPR are calculated by:

$$\text{TPR} = \frac{\# \text{ Positive testing images which are classified as positive}}{\# \text{ Positive testing images}} \times 100\% \quad (4.8)$$

$$\text{FPR} = \frac{\# \text{ Negative testing images which are classified as positive}}{\# \text{ Negative testing images}} \times 100\% \quad (4.9)$$

The ROC curve reflects the TPR and the FPR at the different operating points and the AUC is the area under the ROC curve. It is shown in Fig. 4.11 and Table 4.5 that the proposed two-stage method consistently outperforms the individual methods at most operating points.

Figure 4.11: The ROC curves for the proposed method (the blue solid line), the method of stage 1 (the red dotted line) and the method of stage 2 (the black dashed-dotted line), respectively.

Table 4.5: The AUC, TPR and FPR for the PAP smear dataset

	AUC	TPR	FPR
Method of stage 1	0.67 ± 0.03	$95.41\% \pm 2.0\%$	$39.80\% \pm 2.8\%$
Method of stage 2	0.79 ± 0.04	$92.89\% \pm 2.1\%$	$51.22\% \pm 2.2\%$
The proposed method	0.87 ± 0.04	$97.04\% \pm 1.6\%$	$38.57\% \pm 2.5\%$

Analysis of experiment 1

The complementary effect in experiment 1 validates our analysis of the superiority of the proposed method from three perspectives. The first is the complementary impact from the two-stage structure embedding a reject option. It is designed that only those testing samples which cannot be classified at stage 1 with sufficiently high confidence are to be fed into stage 2.

The second is the complementary impact from the local features. The adoption of SIFT and SURF is complementary since it has been experimentally shown in [50] and [51] that the SIFT is more robust to changes in scale, image blur and rotation while SURF is more robust to changes in illumination.

Thirdly, the BoW models for both stages are complementary in terms of whether spatial information is incorporated. It is observed that some categorical biomedical patterns have high geometric correspondence while some other categorical patterns might have low or no geometric correspondence. Generally, since the geometric correspondence property depends on the category, whether a testing image has or does not have high geometric correspondence is unknown. It is noted that either individual method of each stage lacks the flexibility to handle images with or without geometric correspondence simultaneously. In contrast, the proposed method provides such flexibility based on different formulations of the BoW models for the two stages, respectively. In particular, at stage 1, the spatial pyramid matching (SPM) is applied to incorporate spatial information. At stage 2, the SPM is not applied and therefore it does not incorporate spatial information.

Table 4.6: The comparison between our two-stage method and the method with a direct fusion of the two features

Methods	2D Hela	PAP smear	Hep-2
The proposed method	92.96% $\pm$ 1.3%	90.96% $\pm$ 0.5%	92.97% $\pm$ 1.0%
Direct fusion of two features	87.18% $\pm$ 2.1%	85.20% $\pm$ 1.6%	85.14% $\pm$ 1.3%

It is worthwhile to further discuss the relationship between the classification

structure and the feature properties. The feature adopted at stage 1 incorporates spatial information. Therefore, the classification accuracy at stage 1 is high with respect to those biomedical patterns which have high geometric correspondence. However, stage 1 may misclassify (or classify with lower confidence) those biomedical patterns which have low geometric correspondence. Stage 2 does not consider spatial information but applies ERE to extract discriminative subspace features. It is to classify the images without clear geometric correspondence, but it may fail if the input biomedical patterns have high geometric correspondence. The proposed two-stage classification structure can tackle the two cases involving patterns with or without geometric correspondence simultaneously since only those images with low geometric correspondence are rejected from stage 1 and fed into stage 2. It is noted that a direct fusion of the features from both stages may not improve the classification performance for our application. Since most biomedical images have either high geometric correspondence or not, a direct fusion of the two features at the two stages will generate a potentially confusing representation. Table 4.6 presents that the proposed method outperforms the method with a direct fusion of the two features. In summary, as evidenced in the experimental results, the proposed method integrates the two individual methods of the separate stages.

As an illustration of the complementary effect of the proposed method, Fig. 4.12(a) shows an example of the classification of a biomedical image with high geometric correspondence. The ground truth label of this image is the class ER. It is seen that this pattern is circular with a clear boundary and centroid. This indicates that it has relatively high geometric correspondence. For such type of patterns, the method at stage 1 classifies it with high confidence (as seen from the confidence score defined for each class) and the method at stage 2 misclassifies it as it does not consider spatial information. For the proposed method, since it has been classified at stage 1 with high confidence and will not be fed to stage 2, the proposed method correctly predicts its true label. On the other hand, Fig. 4.12(b) shows another example of a biomedical image without clear geometric correspondence. Its ground truth label is the class Golgi

(a) An example of a biomedical image with high geometric correspondence.

(b) An example of a biomedical image without high geometric correspondence.

Figure 4.12: An illustration of the classification of two biomedical patterns from 2D Hela dataset : (a) an image with high geometric correspondence. (b) an image without high geometric correspondence.

Giantin. It is noted from the confidence score of stage 1 that it cannot be identified with high confidence as the scores for the classes “Golgi Giantin” and “Golgi GPP130” are almost equal. If no reject option is formulated, it would be misclassified as “Golgi GPP130” at stage 1. The method at stage 2 correctly predicts the true label for this image as it avoids being misled by ambiguous spatial information. Using the proposed method, this image is firstly rejected at stage 1 but then correctly classified at stage 2. From these two examples, the complementary effect of the two stages in our proposal is clear.

4.3.4 Experiment 2

In this subsection, we firstly make a comparison between the proposed method and a recently published method which trains an SVM based on an LBP variant called spatial adjacent histogram based on adaptive local binary patterns (SAHLBP) [43]. SAHLBP improves the original LBP by assigning an adaptive neighborhood radius for each pixel and proposes a spatial adjacent histogram strategy to encode the micro-structures for biomedical images. It has been discussed in [43] on three coding schemes conducted for local binary pattern generation. Correspondingly, there are three types of SAHLBP, namely $\text{SAHLBPT}_i (i = 1, 2, 3)$. A linear kernel SVM is applied as discussed in [43].

Table 4.7: Classification accuracy comparison among the proposed method and the SAHLBP based methods

Methods	2D Hela	PAP smear	Hep-2
$\text{SAHLBPT}_1 + \text{SVM}$	$89.68\% \pm 2.0\%$	$86.69\% \pm 2.1\%$	$91.89\% \pm 0.8\%$
$\text{SAHLBPT}_2 + \text{SVM}$	$87.29\% \pm 1.6\%$	$84.03\% \pm 2.1\%$	$88.59\% \pm 0.8\%$
$\text{SAHLBPT}_3 + \text{SVM}$	$90.06\% \pm 1.5\%$	$88.03\% \pm 1.7\%$	$91.86\% \pm 0.9\%$
The proposed method	$92.96\% \pm 1.3\%$	$90.96\% \pm 0.5\%$	$92.97\% \pm 1.0\%$

Table 4.7 shows the classification accuracies of the proposed method and the SAHLBP based methods [43]. From the table, it is noted that in general, the proposed method consistently outperforms the SAHLBP based methods in accuracy for all the datasets. To further compare the classification performance for different categories of biomedical images, we also tabulate the confusion matrices for the two multi-categorical

C1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C3	0.00	0.05	0.89	0.00	0.00	0.00	0.06	0.00	0.00	0.00
C4	0.00	0.05	0.06	0.89	0.00	0.00	0.00	0.00	0.00	0.00
C5	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
C6	0.00	0.00	0.00	0.00	0.06	0.94	0.00	0.00	0.00	0.00
C7	0.00	0.00	0.06	0.00	0.00	0.00	0.88	0.00	0.00	0.06
C8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
C9	0.00	0.00	0.07	0.13	0.00	0.00	0.00	0.00	0.80	0.00
C10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10

(a)

C1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C3	0.00	0.00	0.72	0.00	0.17	0.00	0.11	0.00	0.00	0.00
C4	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.06	0.00
C5	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
C6	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
C7	0.00	0.00	0.12	0.00	0.00	0.00	0.88	0.00	0.00	0.00
C8	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.94	0.00	0.00
C9	0.00	0.00	0.07	0.07	0.00	0.07	0.00	0.00	0.79	0.00
C10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10

(b)

Figure 4.13: The confusion matrices for the 2D Hela dataset generated by (a) the proposed method; (b) the best SAHLBP based method. C1: Actin; C2: DNA; C3: Endosome; C4: Er; C5: Golgia; C6: Golgpp; C7: Lysosome; C8: Mircrotubules; C9: Mitochondiria; C10: Nucleolus.

datasets, i.e., 2D Hela and Hep-2. For each confusion matrix, the diagonal entry indicates the average classification accuracy over 30 rounds of experiments for the corresponding category. The entry at the i_{th} row and the j_{th} column indicates the fraction of the images with ground truth category i but being misclassified as category j . We compare the confusion matrices generated by the proposed method with those benchmark confusion matrices reported in [43] as shown in Figs. 4.13 and 4.14. For the 2D Hela dataset, compared with the benchmark confusion matrix (Fig. 4.13b), the proposed method achieves better performance for the categories of the endosome, the microtubules and the mitochondria while produces lower classification accuracies for the categories of the *er* and the *golpp*. Our method produces perfect classification for the categories of the Actin, the DNA, the Golgia, the microtubules and the nucleolus. For the Hep-2 dataset, compared with the benchmark confusion matrix (Fig. 4.14b), the proposed method produces overall better performance for the categories of the centromere, the coarse speckled, the cytoplasmatic and the homogeneous while with comparable performance for the categories of the fine speckled and Nucleolar. To summarize, the proposed method improves the classification accuracies for most categories compared with the competing methods proposed in [43].

Secondly, we make a comparison between the proposed method and several competing methods under the bag-of-words (BoW) framework which are reported in [43]. The competing methods adopt different local features such as standard SIFT, LBP, SAHLBP and a combination of SIFT and SAHLBP [43]. To obtain a fair comparison, we follow the settings in [43]. The patch size for SIFT is 16 pixels and the grid space is set at 8 pixels. The threshold set for the standard LBP is the central pixel value and that for SAHLBP is set at the mean gray value of the local blocks. These features are encoded using vector quantization (VQ) and SPM is applied with the same parameters as our proposed method. The combination of SIFT and SAHLBP uses both local features to generate the visual words using K-means clustering (named SIFT+SAHLBP for short [43]). For the two random forest based methods using our bag-of-words features

C1	0.97	0.01	0.00	0.00	0.02	0.00
C2	0.05	0.95	0.00	0.00	0.00	0.00
C3	0.00	0.00	1.00	0.00	0.00	0.00
C4	0.05	0.05	0.00	0.76	0.12	0.02
C5	0.00	0.00	0.00	0.02	0.98	0.00
C6	0.00	0.00	0.00	0.00	0.00	1.00
	C1	C2	C3	C4	C5	C6

(a)

C1	0.96	0.00	0.00	0.03	0.01	0.00
C2	0.00	0.90	0.00	0.10	0.00	0.00
C3	0.04	0.10	0.86	0.00	0.00	0.00
C4	0.05	0.03	0.00	0.80	0.12	0.00
C5	0.00	0.02	0.00	0.00	0.98	0.00
C6	0.00	0.00	0.00	0.00	0.00	1.00
	C1	C2	C3	C4	C5	C6

(b)

Figure 4.14: The confusion matrices for the Hep-2 cell dataset generated by (a) the proposed method; (b) the best SAHLBP based method. C1: Centromere; C2: Coarse speckled; C3: Cytoplasmatic; C4: Fine speckled; C5: Homogeneous; C6: Nucleolar.

at the two stages, the number of trees of the random forest model is tuned through cross validation to produce the best generalization performance on the validation set [82].

Table 4.8 shows the average classification accuracies of the compared methods. It is

Table 4.8: Classification accuracy comparison among the proposed method and the competing methods under the bag-of-words framework

Methods	2D Hela	PAP smear	Hep-2
SIFT(BoW+SPM+SVM) [49]	83.79% $\pm$ 2.5%	84.03% $\pm$ 2.3%	80.88% $\pm$ 1.5%
LBP(BoW+SPM+SVM) [48]	81.47% $\pm$ 2.1%	81.43% $\pm$ 2.1%	77.52% $\pm$ 1.9%
SAHLBP(BoW+SPM+SVM) [43]	84.49% $\pm$ 2.2%	86.21% $\pm$ 2.0%	85.06% $\pm$ 2.0%
SIFT+SAHLBP(BoW+SPM+SVM) [43]	86.20% $\pm$ 2.5%	87.63% $\pm$ 2.1%	83.47% $\pm$ 1.8%
SIFT(BoW+SPM+Random forest) [82]	87.38% $\pm$ 1.6%	84.49% $\pm$ 1.8%	85.98% $\pm$ 1.2%
SURF(BoW+ERE+Random forest) [82]	83.75% $\pm$ 1.6%	79.46% $\pm$ 1.9%	82.47% $\pm$ 1.1%
The proposed method	92.96% $\pm$ 1.3%	90.96% $\pm$ 0.5%	92.97% $\pm$ 1.0%

observed that both the SVM and the random forest classifiers achieve the comparable performance using a single type of features. Within the SVM based methods, the SIFT+SAHLBP generally outperforms other competing methods using only one local feature for the 2D Hela and the PAP smear datasets. For the Hep-2 dataset, the SAHLBP outperforms the SIFT+SAHLBP in terms of accuracy. This shows that a direct combination of two types of local features may provide misleading information for this dataset. By comparison, the proposed method significantly outperforms all the competing methods in terms of classification accuracy. These results show that our proposed two-stage method effectively fuses the two types of features with the classification structure which contributes to improving the classification accuracy.

4.4 Summary

In this chapter, we have proposed a two-stage method for robust biomedical image classification. For a testing image, it first goes through an SVM with a reject option at stage 1. It was either classified with high confidence or rejected and then fed into the next stage. At stage 2, we applied eigenfeature regularization and extraction (ERE) to classify the rejected samples using a second SVM trained in the transformed

subspace. The novelties of the proposed method are summarized as: (i) we apply a novel confidence measure which combines the signed distance from the SVM hyperplanes (the discriminative score) and the distance to each categorical centroid (the prior score). It is shown that the proposed score can generally better measure the confidence level of biomedical images than the simple signed distances to SVM hyperplanes; (ii) we propose a novel approach to determining the threshold for the confidence scores using an SVM with a reject option based on Chow's rule at stage 1; (iii) at stage 2, we apply eigenfeature regularization and extraction (ERE) to learn a discriminative subspace to classify the rejected images from stage 1; (iv) the method is specifically designed to solve biomedical image classification by building various BoW features with and without incorporating spatial information for the two stages, respectively. Using three benchmark datasets, the proposed method significantly outperformed several competing state-of-the-art methods in terms of multiple classification performance metrics.

The purpose of the SVM with a reject option formulated in this chapter is to select ambiguous or outlier samples from Stage 1 and feed to Stage 2 for further classification. Since Stage 1 is designed to classify the biomedical images with high geometric correspondence, the rejected samples from Stage 1 could be those biomedical images with low geometric correspondence. From the analysis in Section 4.3.3, the rejected samples with low geometric correspondence could be classified in the discriminative subspace learned by ERE. Hence, the proposed two-stage method is specifically designed to handle the ambiguous and outlier biomedical patterns with different geometric correspondence.

Chapter 5

A two-stage method for robust ring-like endosome detection

Chapter 5 proposes a two-stage method for robust biomedical image detection. We investigate a specific task of detecting ring-like endosomes in fluorescent microscopy images. In Section 5.1, we introduce the background of ring-like endosome detection and point out the problems encountered by the existing methods. Then, the specifications of the proposed method are presented in Section 5.2. Section 5.3 presents and analyzes the experimental results on real-world microscopy image datasets. Section 5.4 concludes this chapter.

5.1 Introduction

In this chapter, we investigate robust methods for biomedical image detection. Specifically, we focus on a task of detecting ring-like endosomes in fluorescent microscopy images. Fluorescent microscopy is a modern technology which can produce snapshots of subcellular structures inside cells (e.g., see Fig. 5.1a). Endosomes (shown in Fig. 5.1b) are organelles which can be found in all eukaryotic cells and function as important transport compartments that shuttle proteins, nutrients and other materials inside cells [84]. The detection of ring-like endosomes from non-endosome background patterns (e.g., see Fig. 5.1c) is of significant biological interest related to the analysis of interactions among proteins and organelles. For example, ring-like endosomes are closely relevant to the assessment of the effectiveness of a particular class of drugs called therapeutic monoclonal antibodies [85]. The annotation of endosomes is usually performed manually, which is time-consuming and inaccurate if carried out by inexperienced analysts. Hence, it is valuable to automate the process of endosome

Figure 5.1: (a) A microscopy image of a cell with multiple subcellular structures. (b) Three patches of ring-like endosomes. (c) Three patches of non-endosome background patterns. The scale bar for (a) is $5\ \mu\text{m}$ and that for (b) and (c) is $1\ \mu\text{m}$.

Figure 5.2: A microscopy image with manual annotations (using yellow rectangles) of ring-like endosomes in a microscopy image.

detection. Fig. 5.2 illustrates the manual annotations of ring-like endosomes in a microscopy image.

In Section 2.4.2, we have reviewed both unsupervised and one-stage supervised detection methods for biomedical images. In general, most unsupervised methods are computationally efficient and perform well when the targets have simple and distinct features, e.g., detecting circular targets from highly non-circular background patterns. However, unsupervised methods normally fail to detect complicated targets especially when the adopted prior information is not sufficiently discriminative to distinguish the targets from some background patterns. In such scenarios, supervised methods generally outperform unsupervised methods as they apply the discriminative information between the targets and the non-target background patterns. Therefore, in this chapter, we focus on improving supervised methods.

A disadvantage of one-stage supervised methods is the adoption of the sliding window search as it is generally computationally complicated and might miss the detection if the target is partially covered or “mismatched” with the sliding window, e.g., a small target stays inside a relatively large sliding window. Also, the most important factor for a one-stage supervised method is the performance of its classifier. Such performance is highly dependent on the selection of training patches as well as whether each training patch is represented by a discriminative feature vector. Most one-

stage supervised methods only exploit global information of each training patch, i.e., representing a training patch (as a whole) by a feature vector, which might deteriorate the detection performance.

Motivated by the above discussions, we propose a two-stage supervised method for the automated detection of ring-like endosomes in fluorescent microscopy images. The proposed method consists of a localization stage cascaded by an identification stage. Given a testing microscopy image, the localization stage generates a voting-map by locally comparing the selected query patches and the testing image based on a bag-of-words model. Using the voting-map, a number of candidate patches of endosomes are determined. Subsequently, in the identification stage, a support vector machine (SVM) is trained from the training patches represented by a locality-constrained linear coding (LLC) based bag-of-words model. Each of the candidate patches is classified by the SVM to rule out those patches of endosome-like background patterns.

The proposed method alleviates the problems encountered by those one-stage supervised methods. In the localization stage, the proposed method directly localizes potential candidate patches of endosomes using the voting-map based approach and thereby avoids the computationally complicated sliding window search. Each candidate patch locates a potential endosome at its center and the size of the patch can be automatically determined, which alleviates the “mismatched” effects caused by the sliding window search. Also, unlike one-stage methods which only exploit global information of training patches, the proposed method applies the local information of training patches in the localization stage and the global information in the identification stage. Both stages perform complementarily to improve the detection performance.

The performance of the proposed method is evaluated by the experiments on real-world microscopy images of human myeloid endothelial cells. It is shown that the proposed method significantly outperforms several state-of-the-art competing methods using multiple performance metrics.

5.2 Methodology

The proposed method consists of two stages: localization and identification. The localization stage is designed to provide location information of candidate patches and the identification stage is to further identify endosomes in those candidate patches from a classification perspective.

5.2.1 Feature extraction

In this chapter, we extract dense SIFT features from microscopy images and build bag-of-words models for image representation. The SIFT feature and the bag-of-words model have been reviewed in Section 2.3.1. Note that even though both the methods proposed in the two stages are formulated based on this feature extraction scheme, the bag-of-words models are exploited differently for separate stages. For the localization stage, we compare and find matched SIFT feature pairs between the query patches and the testing image to determine potential locations of endosomes. For the identification stage, to train an SVM to classify endosomes from non-endosomes, we exploit the LLC-based bag-of-words model (reviewed in Section 2.3.1) to summarize SIFT features for discriminative image representation. The following subsections present the details of the localization and the identification stages, respectively.

5.2.2 Localization stage

The localization stage is proposed to determine candidate patches of endosomes in a microscopy image. Since endosomes are approximately circular with ring-like textures, a candidate patch of an endosome in a microscopy image can be determined by the centroid of the endosome (a pixel in the microscopy image) and a rectangle box which contains the endosome as shown in Fig. 5.3a. To find the centroids of the candidate endosomes, a voting-map, i.e., an image of the same size as the microscopy image and each pixel of the voting-map quantifies the confidence of an endosome centroid,

Figure 5.3: (a) The candidate patch of an endosome in a microscopy image can be determined by the centroid of the endosome (denoted by the red dot) and a rectangle box which contains the whole endosome (denoted by dashed green rectangles). (b) The voting-map for the microscopy image shown in (a). Scale bar = 1 μm .

is generated in the localization stage. Fig. 5.3b shows an example of the voting-map associated with the microscopy image in Fig. 5.3a.

Motivated by the techniques proposed in [54] on natural object localization using one query patch of the object, we propose an improved category localization method for our application using multiple query patches of endosomes. Here, a query patch of an endosome refers to a patch of a standard endosome selected by hands. In this stage, each of the query patches and the testing microscopy image are locally compared to find the potential locations of endosome centroids and “votes” on these locations by a quantity called discriminative capability [86]. After performing such a measurement between every query patch and the testing microscopy image, a voting-map is generated. The remaining parts of this subsection show the technical details of the localization stage consisting of a training phase and a testing phase.

Training phase

In the training phase, a set of visual words are first generated from the selected query and background patches. Then a bag-of-words histogram is calculated for each patch. The location and assignment (to certain visual word) for each SIFT are recorded. Finally, the discriminative capability of each visual word is determined. These quantities will be

adopted in the testing phase for voting-map generation.

Specifically, we manually collect M (e.g., $M = 50$) patches of endosomes as a query set denoted by $\mathbf{Q}_{set} = \{\mathbf{Q}_1, \mathbf{Q}_2, \dots, \mathbf{Q}_M\}$ and another set of N (e.g., $N = 200$) patches of background patterns denoted by $\mathbf{B}_{set} = \{\mathbf{B}_1, \mathbf{B}_2, \dots, \mathbf{B}_N\}$. The size of each of these patches is set at 45×45 pixels. The local SIFT features are extracted from each of the patches in $\mathbf{Q}_{set}$ and $\mathbf{B}_{set}$. Then a bag-of-words dictionary with K (e.g., $K = 1000$) visual words is generated from these SIFT features. With the dictionary, the bag-of-words histograms are calculated for the patches. For each query patch in $\mathbf{Q}_{set}$, the location of each of its SIFT features, i.e., the center pixel position of the grid where the SIFT is calculated, is recorded.

Subsequently, the discriminative capability [86] is calculated to quantify the significance of each visual word in discriminating the patches of endosomes from those of non-endosome background patterns. The discriminative capability of a visual word is defined as the ratio between the within-class similarity and the between-class similarity [86]. The between-class similarity $\delta(k)$, $k = 1, 2, \dots, K$, for the k_{th} visual word is defined using the bag-of-words histograms of the patches from $\mathbf{Q}_{set}$ and $\mathbf{B}_{set}$ as:

$$\delta(k) = \frac{1}{M \cdot N} \sum_{m=1}^M \sum_{n=1}^N \min(\mathbf{h}_{\mathbf{Q}_m}(k), \mathbf{h}_{\mathbf{B}_n}(k)), \quad (5.1)$$

where $\mathbf{h}_{\mathbf{Q}_m}$ and $\mathbf{h}_{\mathbf{B}_n}$ denote the K -dimensional bag-of-words histograms of the m_{th} query patch $\mathbf{Q}_m$, $m = 1, 2, \dots, M$, and the n_{th} background patch $\mathbf{B}_n$, $n = 1, 2, \dots, N$, respectively. The double summation in Eq. (5.1) is over the patch indices m and n for $\mathbf{Q}_{set}$ and $\mathbf{B}_{set}$, respectively. A large $\delta(k)$ shows that the similarity among the patches from different classes is high for the k_{th} visual word, i.e., the k_{th} visual word is significant to classify the patches from different classes. On the other hand, the within-class similarity $\alpha(k)$ associated with the endosome class ($\mathbf{Q}_{set}$) for the k_{th} visual word is defined as:

$$\alpha(k) = \frac{1}{M \cdot (M-1)} \sum_{m=1}^M \sum_{l=1, l \neq m}^M \min(\mathbf{h}_{\mathbf{Q}_m}(k), \mathbf{h}_{\mathbf{Q}_l}(k)). \quad (5.2)$$

A large $\alpha(k)$ indicates that the similarity among the patches from the same class is high for the k_{th} visual word, i.e., the k_{th} visual word is significant to represent the patches from the class. Finally, the discriminative capability $d(k)$ for the k_{th} visual word is defined as:

$$d(k) = \frac{\alpha(k)}{\delta(k)}. \quad (5.3)$$

Calculating Eqs. (5.1) to (5.3) for $k = 1, 2, \dots, K$, the discriminative capability of each visual word is determined. A large $d(k)$ comes from either a small $\delta(k)$ or a large $\alpha(k)$. Therefore, a large $d(k)$ indicates that the k_{th} visual word is effective not only to classify the patches from different classes but also to represent the patches from the same class.

Testing phase

With a testing microscopy image $\mathbf{D}$, the goal is to search for the candidate patches within $\mathbf{D}$ which are visually similar to at least one query patch in $\mathbf{Q}_{set}$. To locate candidate patches, a voting-map based approach is exploited. Each pixel of the voting-map quantifies the confidence of an endosome centroid at the corresponding position of $\mathbf{D}$.

Specifically, the SIFT features are extracted from $\mathbf{D}$ and each of these SIFT features is assigned to a visual word generated in the training phase. To locally compare each query patch and the testing microscopy image $\mathbf{D}$, a SIFT feature from a query patch is defined as ‘matched’ to another SIFT feature from $\mathbf{D}$ if both SIFT features are assigned to the same visual word. Assuming a pair of matched features $\mathbf{f}$ and $\mathbf{g}$ extracted from $\mathbf{Q}_m$ and $\mathbf{D}$ is assigned to the k_{th} visual word, the location of $\mathbf{f}$ and its relative bias vector with respect to the centroid $\mathbf{c}_{\mathbf{Q}_m}$ of the query patch $\mathbf{Q}_m$ can be determined geometrically by $L(\mathbf{f}) - L(\mathbf{c}_{\mathbf{Q}_m})$. $L(\mathbf{f})$ is the center pixel position of the grid where the SIFT feature $\mathbf{f}$ is calculated and $L(\mathbf{c}_{\mathbf{Q}_m})$ is the center pixel position of the query patch $\mathbf{Q}_m$. Subsequently, a possible location of the centroid of a candidate patch of an endosome in $\mathbf{D}$ denoted as $\mathbf{c}_D$ is determined by $L(\mathbf{c}_D) = L(\mathbf{g}) - s * (L(\mathbf{f}) - L(\mathbf{c}_{\mathbf{Q}_m}))$. $L(\mathbf{g})$ is the center pixel

Figure 5.4: A diagram illustrating the process of voting-map generation. Three query patches and two scaling factors, i.e., 0.8 and 1, are adopted in this example. (a) The testing microscopy image. (b) The three query patches. (c) The generation of the voting-maps with the scaling factor set at 0.8. For each query patch, a voting-map is generated by local feature matching between the query patch and the testing image. Then a voting-map associated with the scaling factor of 0.8 is obtained by summing the three voting-maps. (d) The generation of the voting-maps with the scaling factor set at 1. (e) The final voting-map is determined by selecting the optimal scaling factor for each pixel position as described in the text. Scale bar = 1 μm .

position of the grid where the SIFT feature $\mathbf{g}$ is calculated and s is the scaling factor. Since the endosomes appearing in $\mathbf{D}$ could be similar to a scaled version (by a scaling factor s) of the endosome in the query patch, we search through multiple scaling factors, i.e., $s_1, s_2, \dots, s_{n_s}$, within the range of $[0.5, 1.5]$ based on the knowledge of the potential sizes of an endosome in a microscopy image.

After determining $\mathbf{c}_D$, we can calculate the pixel value of the voting-map $\mathbf{V}(s, \mathbf{Q}_m, \mathbf{D})$, which is defined as an image associated with s , $\mathbf{Q}_m$ and $\mathbf{D}$. First, all the pixel values of $\mathbf{V}(s, \mathbf{Q}_m, \mathbf{D})$ are initialized as zero. For a pair of matched features being

assigned to the k_{th} visual word, $\mathbf{V}(s, \mathbf{Q}_m, \mathbf{D})$ is accumulated at the location $\mathbf{c}_D$ with a value of the discriminative capability $\mathbf{d}(k)$ which has been calculated in the training phase. By traversing all the matched feature pairs, the voting-map $\mathbf{V}(s, \mathbf{Q}_m, \mathbf{D})$ is generated. To integrate the information from multiple query patches, a general voting-map $\mathbf{V}(s, \mathbf{Q}_{set}, \mathbf{D})$ associated with s , $\mathbf{Q}_{set}$, and $\mathbf{D}$ can be determined as the summation of the voting-map associated with every query patch in $\mathbf{Q}_{set}$:

$$\mathbf{V}(s, \mathbf{Q}_{set}, \mathbf{D}) = \sum_{m=1}^M \mathbf{V}(s, \mathbf{Q}_m, \mathbf{D}). \quad (5.4)$$

To make the notation less clumsy, we simplify the notation for $\mathbf{V}(s, \mathbf{Q}_{set}, \mathbf{D})$ as $\mathbf{V}_s$. $\mathbf{V}_s(x, y)$ denotes the value of the pixel at the x_{th} row and the y_{th} column of $\mathbf{V}_s$. After repeating the same procedure for all the scaling factors $s_1, s_2, \dots, s_{n_s}$, n_s voting-maps $\mathbf{V}_{s_1}, \mathbf{V}_{s_2}, \dots, \mathbf{V}_{s_{n_s}}$ are generated. A final voting-map $\mathbf{V}_{final}$ is determined by selecting the optimal scaling factor for each pixel position. Specifically, for the position (x_v, y_v) , an optimal scale $s_{(x_v, y_v)}^*$ is determined by selecting the scaling factor which produces the maximum pixel value of $\mathbf{V}_{s_1}, \mathbf{V}_{s_2}, \dots, \mathbf{V}_{s_{n_s}}$ at (x_v, y_v) , i.e.,:

$$s_{(x_v, y_v)}^* = \arg \max_{s \in \{s_1, s_2, \dots, s_{n_s}\}} \mathbf{V}_s(x_v, y_v). \quad (5.5)$$

The value of the final voting-map $\mathbf{V}_{final}$ at (x_v, y_v) is determined as:

$$\mathbf{V}_{final}(x_v, y_v) \triangleq \mathbf{V}_{s_{(x_v, y_v)}^*}(x_v, y_v). \quad (5.6)$$

The final voting-map $\mathbf{V}_{final}$ and the corresponding $s_{(x_v, y_v)}^*$ for each pixel in $\mathbf{V}_{final}$ will be adopted in the identification stage. Fig. 5.4 illustrates the voting-map generation procedure.

It is worthwhile to further discuss the “spatial consistency” of the voting during the local feature matching process. Fig. 5.5 illustrates the details of local feature matching in the voting-map generation process. In this figure, given a query patch, three local

Figure 5.5: A diagram illustrating two cases of local feature matching in the voting-map generation procedure.

features (denoted by the black triangle, circle, and diamond, respectively) and their relative bias vectors with respect to the centroid of the query patch are shown. The matched local features are denoted by the same symbol. Here we consider two specific cases which are shown in the dashed rectangles $S1$ and $S2$, respectively. In $S1$, three matched features are spatially consistent with their counterparts in the query patch. In this case, three features vote to a consistent location. Whereas in $S2$, three matched features are spatially inconsistent with their counterparts and therefore the locations to which they vote are not geometrically close. The spatially inconsistent features are not useful for localization. Since the locations voted by spatially consistent features have much higher voting-map pixel values compared with those voted by the inconsistent features based on the experimental observations, we can simply threshold the voting-map to rule out the locations voted by spatially inconsistent features as described in the subsequent section.

5.2.3 Identification stage

After the localization stage, we determine the voting-map $\mathbf{V}_{final}$ and the optimal scaling factor corresponding to each pixel in $\mathbf{V}_{final}$. Candidate patches of endosomes centered at the locations with high voting-map pixel values can be obtained from the testing

microscopy image. The identification stage is proposed to train a support vector machine to identify endosome patches and rule out endosome-like background patches from the candidate patches provided by the localization stage. The identification stage also consists of a training phase and a testing phase.

Training phase

In the training phase, we train a binary support vector machine (SVM) to identify endosomes from the candidate patches. From a classification perspective, we take endosomes as positive class and non-endosome background patterns as a negative class. We adopt the same query and background patches used in the localization stage as the positive and negative training patches, respectively.

To represent each training patch, we apply a modified BoW model where locality-constrained linear coding (LLC) [53] is adopted for local feature assignment. The details of LLC have been reviewed in Section 2.3.1. A binary SVM with a linear kernel is trained to classify the training patches represented by the bag-of-words model. The details of determining the optimal SVM hyperplane by solving a quadratic programming problem can be seen in [37]. The regularization parameter C of the SVM is selected through cross validation [37].

Testing phase

In the testing phase, We first normalize the voting-map to the range $[0, 1]$ and set a threshold γ for the voting-map $\mathbf{V}_{final}$ to rule out those locations with very low confidence of endosome centroids. Consequently, the thresholded voting-map only retains the pixels whose values are larger than γ and sets the other pixels to be zero. Subsequently, the candidate patches of endosomes are determined. Specifically, for a non-zero pixel location $\{x_o, y_o\}$ in the thresholded voting-map, the optimal scale $s_{(x_o, y_o)}^*$ which is recorded in the localization stage is retrieved. Assuming the width and height of each query patch being w and h , respectively, a candidate patch can be obtained and

parameterized as $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$, where $(x_{\min}, y_{\min})$ and $(x_{\max}, y_{\max})$ denote the pixel positions of the upper-left and the bottom-right corners of the candidate patch, respectively. They can be calculated as:

$$\begin{aligned} x_{\min} &= \left\lceil x_o - s_{(x_o, y_o)}^* \times 0.5 \times w \right\rceil, \\ y_{\min} &= \left\lceil y_o - s_{(x_o, y_o)}^* \times 0.5 \times h \right\rceil, \\ x_{\max} &= \left\lceil x_o + s_{(x_o, y_o)}^* \times 0.5 \times w \right\rceil, \\ y_{\max} &= \left\lceil y_o + s_{(x_o, y_o)}^* \times 0.5 \times h \right\rceil, \end{aligned} \tag{5.7}$$

where $\lceil \cdot \rceil$ is the ceiling function which takes an input of a real number and outputs the smallest integer which is larger than or equal to the input number. $\lfloor \cdot \rfloor$ is the floor function which outputs the greatest integer which is smaller than or equal to the input number. In this way, a candidate patch $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ and its voting-map value $\mathbf{V}_{final}(x_o, y_o)$ is retrieved. Then, the non-maximum suppression (NMS) algorithm [87] is applied to handle the duplicated detections, i.e., for multiple candidate patches containing the same target, the patch with the highest voting-map value is selected and the others are removed (suppressed). Finally, the remaining candidate patches are classified by the SVM from the training phase as shown in Fig. 5.6.

Algorithm 4 summarizes the overall procedure for the detection of ring-like endosomes from a testing microscopy image. The algorithm consists of the procedure for both the localization stage and identification stage, respectively. The testing microscopy image is fed to the two stages in sequence and the detections of endosomes are automatically determined.

5.3 Experiments

In this section, we first describe the dataset in Subsection 5.3.1. In Subsection 5.3.2, we introduce several metrics to evaluate the detection performance. The experimental protocol and parameter settings for the method comparison are specified in Subsection

Algorithm 4: The procedure for the detection of endosomes from a testing image.

Input:

D : the testing image.

Q_m : the m_{th} query patch in the query set Q_{set} .

M : the number of query patches.

K : the number of visual words.

n_s : the number of the elements in the scaling factor set

$s = \{s_1, s_2, \dots, s_{n_s}\}$.

1 **Initialization:** Initialize the pixel values for the voting-maps

$V_{s_1}, V_{s_2}, \dots, V_{s_{n_s}}$ associated with each scaling factor as zeros.

2 **Localization Stage:**

3 **for** $s = s_1$ to s_{n_s} **do**

4 **for** visual word $k = 1$ to K **do**

5 **for** query patch $m = 1$ to M **do**

6 **for** local features $f \in Q_m$ and $g \in D$ **do**

7 **if** f and g are both assigned to the k_{th} visual word **then**

8 $L(c_D) = L(g) - s * (L(f) - L(c_{Q_m}))$.

9 $V_s(L(c_D)) = V_s(L(c_D)) + d(k)$.

10 **for** Each location $\{x_v, y_v\}$ **do**

11 Determine the optimal $s_{(x_v, y_v)}^*$ by (5.5).

12 Determine the values of the final voting-map V_{final} by (5.6).

13 **Identification Stage:**

14 Normalize the values of the final voting-map V_{final} into range $[0, 1]$ and threshold the final voting-map by a threshold γ .

15 **for** Each location $\{x_o, y_o\}$ with non-zero thresholded voting-map value **do**

16 The optimal scale $s_{(x_o, y_o)}^*$ and $V_{final}(x_o, y_o)$ are retrieved.

17 A candidate patch centered at $\{x_o, y_o\}$ is determined by (5.7).

18 Apply the non-maximum suppression (NMS) algorithm to remove the duplicated detection patches.

19 **for** Each candidate patch P **do**

20 Calculate the LLC-based bag-of-words histogram from P .

21 Apply the SVM to classify the histogram and determine if P contains an endosome or not.

Figure 5.6: A diagram illustrating the identification stage using SVM. (a) After the voting-map thresholding and non-maximum suppression, the candidate patches of endosomes are obtained. (b) Each candidate patch is represented by an LLC-based bag-of-words histogram. (c) The histogram is classified by the SVM to determine whether the patch contains an endosome or a non-endosome background pattern. Scale bar = 1 μm .

5.3.3. The experimental results and analyses are presented in Subsection 5.3.4.

The competing methods include two one-stage supervised learning methods and one unsupervised learning method. For the supervised learning methods, one is the AdaBoost classifier trained on Haar-like features (AB-Haar). This method was originally proposed in [72] to systematically build a face detector and extended for the detection of molecular particles in live cells in [73]. The second is a linear discriminant analysis (LDA) model trained on simple pixel features (LDA-Pixel) which was proposed in [66] for spot detection within microscopy images. For the competing unsupervised method, we select the method proposed in [70] for cell nuclei detection mainly based on image morphological operations and Fast Radial Symmetric Transform (MO-FRST).

5.3.1 Dataset description

We evaluate the detection performance of the compared methods using the real-world microscopy images of human myeloid endothelial cells. To generate the dataset, the human myeloid endothelial cells were maintained, transfected with DNA plasmids that express green fluorescent protein-tagged neonatal Fc receptor (FcRn-GFP), plated in glass-bottom dishes and imaged as described in [85, 88]. The images were acquired using a Zeiss AxioObserver.Z1 microscopy with widefield arc lamp illumination, a 63X 1.4NA Plan Apochromat oil immersion objective, a GFP-specific filterset (GFP-3035D-000, Semrock) and a monochromatic CCD camera (Orca ER, Hamamatsu). The dataset consists of two groups of microscopy images with different total magnifications, i.e., 100.8X and 157.5X, respectively. In total, there are 50 microscopy images which contain 1346 endosomes annotated by the analysts. Table 5.1 summarizes the specific information of the dataset.

Table 5.1: The dataset information

Total magnification	100.8X	157.5X
# of microscopy images	30	20
# of annotated endosomes	765	581

5.3.2 Performance metrics

The detection performance is evaluated based on an overlay of the detected patches and the ground truth patches annotated by the analysts. A detected patch is defined as correct if the area of overlap (α_0) between the detected patch b_p and the ground truth patch b_{gt} exceeds 50%, where α_0 is defined as

$$\alpha_0 = \frac{\text{area}(b_p \cap b_{gt})}{\text{area}(b_p \cup b_{gt})}, \quad (5.8)$$

where $b_p \cap b_{gt}$ denotes the intersection of the detected and ground truth patches and $b_p \cup b_{gt}$ denotes their union [89]. Based on this criterion, three basic performance statistics, namely N_{FN} , N_{TP} and N_{FP} are determined. N_{FN} is the number of false negative detections, i.e., the number of patches which are in the ground truth but not detected by the method. N_{TP} is the number of true positive detections, i.e., the number of patches which are in the ground truth and also detected by the method. N_{FP} is the number of false positive detections, i.e., the number of patches which are detected by the method but not shown in the ground truth.

Using these three statistics, several performance metrics are defined. For classifier based detection methods, the confidence of each detected patch is quantified by a score output from the classifier, e.g., posterior probabilities. By setting thresholds to the confidence scores, we can plot (using log-log plots) the miss rate (MR) against the false positives per image (FPPI) curves [90] [91] to show the overall detection performance, where the miss rate (MR) is calculated by $MR = \frac{N_{FN}}{N_{FN} + N_{TP}}$ and the false positives per image (FPPI) is the average value of N_{FP} over the testing microscopy images. Specifically, given a threshold, a pair of values of the false positives per image (FPPI) and the corresponding miss rate (MR) are determined. By varying the threshold, an MR-FPPI curve can be plotted. A better detection method produces a lower miss rate and a smaller number of the false positive detections and therefore its MR-FPPI curve stays closer to the left-bottom corner [90].

Quantitatively, we calculate four performance metrics, namely the log-average miss rate, the precision, the recall and the F_1 score. The log-average miss rate is defined based on the MR-FPPI curve. It is calculated by averaging miss rate at nine FPPI rates evenly spaced in log-space in the range 10^{-2} to 10^0 [90]. The definitions of the precision (P), the recall (R) and the F_1 score are as follows:

$$P = \frac{N_{TP}}{N_{FP} + N_{TP}}, R = \frac{N_{TP}}{N_{FN} + N_{TP}}, F_1 = \frac{2 \times P \times R}{P + R}. \quad (5.9)$$

The precision is the ratio of the number of the correctly detected patches to that of all the detected patches. The recall is the ratio of the number of the correctly detected patches to that of all the ground truth patches. The F_1 score combines the information from both the precision and the recall. The values of log-average miss rate, the precision, the recall and the F_1 score are within $[0, 1]$. A better detection method produces higher values for the precision, the recall and the F_1 score while a lower value for the log-average miss rate.

5.3.3 Experimental protocol and parameter settings

We conduct 500 rounds of experiments to evaluate the detection performance of the compared methods. For each round, we randomly select 20% of the microscopy images for training and the remaining 80% of the images for testing. This experimental protocol is consistently conducted for the two groups of microscopy images with different total magnifications, respectively. All the performance metrics (including the curves and quantitative results) are calculated as the average (with standard deviations) of the 500 rounds of experiments.

For the proposed method, in the localization stage, we select 50 query patches together with 200 patches of the background patterns to determine the voting-map. The size of these patches is set at 45×45 (pixels). For bag-of-words model generation, each patch (or each testing microscopy image) is divided into 16×16 (pixels) overlapping grids and the grid space of two neighboring grids is set at 2 pixels. A SIFT feature is extracted from each grid. The number of visual words for the bag-of-words dictionary is set at 400. The range for the scaling factor searching is $[0.5, 0.6, \dots, 1.5]$ and therefore n_s is set at 11. The threshold γ for the normalized voting-map is set at 0.2. For locality-constrained linear coding (LLC) [53], the number of visual words selected to encode a SIFT feature is set at 5. For SVM training, the regularization parameter C of the SVM is selected through a 5-fold cross validation within the range of $[2^{-8}, 2^{-7}, \dots, 2^7, 2^8]$.

For the two competing supervised methods, namely, AB-Haar [66, 73] and LDA-

Pixel [66], the positive and negative patches are the same as those adopted in the proposed method. For the AB-Haar method, each training patch is represented by the Haar-like features adopted in [66]. For the LDA-Pixel method, each training patch is represented by a feature vector generated through concatenating the horizontal rows of pixels [66]. The sliding window size is set at 45×45 (pixels) and the two neighboring sliding windows have 14 pixels in overlapping for both horizontal and vertical directions.

For the unsupervised MO-FRST [70], the searching range of the radius of FRST transformation is set at $[16, 18, \dots, 30]$ (pixels) to cover the possible radius of endosomes. The radial strictness parameter in FRST is set at 2, the scaling factor is set at $k_p = 10$ and the size of the pixels for the non-maximum suppression (NMS) is set at 20 [70].

5.3.4 Experimental results and analysis

Figs. 5.7 and 5.8 show the qualitative detection performance for two testing microscopy images with different total magnifications. As shown in Figs. 5.7a and 5.8a, the proposed method approximately localizes each endosome at the center of the detected patch and automatically determines the size of each detected patch. However, Both AB-Haar and LDA-Pixel methods suffer from the “mismatch” between the sliding window and the endosome (see Figs. 5.7b, 5.7c, 5.8b and 5.8c). The MO-FRST can localize the centroid of all the circular structures but fail to further distinguish the endosomes and background patterns which are also circularly shaped (see in Figs 5.7d and 5.8d). The proposed method produces the highest number of true positive detections while the lowest number of false positive and false negative detections compared with the competing methods. It is also noted that the proposed method can better handle the detection of the endosomes which stay very close or even touching to each other since the proposed localization stage locally compares the SIFT features. Hence, even when an endosome only partially appears, there are still some local features being determined as matched to the local features extracted from the query patches. However, it is observed that the performance of the competing methods significantly deteriorates in this scenario.

(a)

(b)

(c)

(d)

Figure 5.7: The detection results for an example microscopy image with the total magnification 100.8X using: (a) the proposed method, (b) AB-Haar, (c) Pixel-LDA, (d) MO-FRST. The true positives, the false positives and the false negatives are represented by green, red, and yellow patches, respectively in the color version of this thesis. In (a), we show three enlarged patches which are a false positive (indicated by a red arrow), a true positive (green arrow) and a false negative (yellow arrow). The scale bars for the big microscopy images is $5 \mu\text{m}$ and that for the three small patches shown in (a) is $1 \mu\text{m}$.

(a)

(b)

(c)

(d)

Figure 5.8: The detection results for another example microscopy image with the total magnification 157.5X using: (a) the proposed method, (b) AB-Haar, (c) Pixel-LDA, (d) MO-FRST. Here, the notations are the same as Fig. 5.7. Scale bar = 5 μm .

Quantitatively, the detection performance of the compared methods is shown by the MR-FPPI curves and the four metrics introduced in Section 5.3.2. Figs. 5.9a and 5.9b show the MR-FPPI curves of the compared methods for the two groups of microscopy images with different total magnifications, respectively. It is observed that the proposed method generally outperforms the competing methods as its MR-FPPI curve stays the closest to the left-bottom corner. Among the competing methods, AB-Haar produces the worst performance and both LDA-Pixel and MO-FRST produce a comparable performance. Table 5.2 presents the quantitative performance metrics of the compared methods (the results for the best method are bolded). It is observed from the table that the proposed method significantly outperforms all the competing methods in terms of producing the lowest log-average miss rate and the highest precision, recall and F_1 score. This indicates the proposed method produces more accurate detections while lower missing detections of endosomes compared with the competing methods.

Table 5.2: Average performance metrics (and standard deviations) for the compared methods

Methods	Log-average miss rate	Precision	Recall	F_1 score
The images with total magnification of 100.8X				
AB-Haar [73]	0.529 (0.006)	0.296 (0.012)	0.660 (0.007)	0.409 (0.011)
LDA-Pixel [66]	0.344 (0.005)	0.615 (0.025)	0.636 (0.008)	0.625 (0.010)
MO-FRST [70]	0.269 (0.009)	0.519 (0.011)	0.761 (0.005)	0.617 (0.007)
Proposed method	0.138 (0.007)	0.812 (0.009)	0.807 (0.005)	0.773 (0.006)
The images with total magnification of 157.5X				
AB-Haar [73]	0.457 (0.022)	0.338 (0.009)	0.695 (0.007)	0.455 (0.008)
LDA-Pixel [66]	0.235 (0.023)	0.475 (0.013)	0.846 (0.009)	0.608 (0.011)
MO-FRST [70]	0.241 (0.015)	0.441 (0.018)	0.833 (0.005)	0.577 (0.016)
Proposed method	0.097 (0.013)	0.819 (0.012)	0.850 (0.015)	0.770 (0.009)

For the comparison with those one-stage supervised methods, since ring-like endosomes are not just simple “bright structures” in the dark background like spots, the features like Haar-like features or pixel intensities adopted for bright spot detection might not be sufficiently discriminative to describe the micro-structures of endosomes. The proposed method extracts SIFT features summarizing local gradients which can robustly capture the distinct local structures of endosomes with the influences of noise and illuminations in the microscopy images. Therefore, the proposed method

(a)

(b)

Figure 5.9: The MR-FPPI curves of the compared methods for (a) the microscopy images with the total magnification of 100.8X, (b) the microscopy images with the total magnification of 157.5X.

outperforms the two one-stage supervised methods by extracting more discriminative features. Moreover, the proposed two-stage detection structure utilizes both local and global information from the endosome patches and the background pattern patches. Specifically, on the one hand, the comparison of SIFT features in the localization stage provides the location of each candidate endosome. On the other hand, the adoption of the LLC-based bag-of-words model to represent the patches for SVM training applies the global information of each training patch as a whole to identify endosomes. Hence, the proposed architecture outperforms the one-stage supervised detection methods whose performance highly relies on the classifiers where only the global information is incorporated.

For the comparison with the unsupervised method, MO-FRST [70] was specifically formulated to detect cell nuclei which are circular and symmetrical “bright structures”. However, endosomes are only approximately circular and symmetrical. Moreover, they have ring-like textures. MO-FRST produces many false positive detections since it simply captures all circular and symmetrical patterns but fails to further identify the ring-like structures. In contrast, the proposed method exploits distinct textures of ring-like endosomes in both stages and hence produces much fewer false positive detections

5.4 Summary

In this chapter, we have proposed a two-stage method for robust biomedical image detection, specifically for the task of detecting ring-like endosomes in fluorescent microscopy images. Essentially, the method exploits local SIFT feature matching to locate candidate patches and then identifies each of the candidate patches from a classification perspective. On the one hand, the proposed method is supervised and therefore it does not require much prior knowledge as unsupervised methods do. On the other hand, the proposed two-stage method alleviates the limitations of one-stage supervised methods reviewed in Section 2.4.2. The experiments on the real-world

microscopy images show that the proposed method can produce significantly better detection results than several competing methods in terms of multiple performance metrics.

Chapter 6

Conclusions and future works

In this chapter, Section 6.1 gives a summary of the major works presented in Chapter 3, 4 and 5, respectively. Some future directions of these works are pointed out in Section 6.2.

6.1 Conclusions

In this thesis, robust classification and detection methods with applications in biomedical images have been studied. The study is motivated by Chow's theoretical works on classification with a reject option and specific applications in biomedical images. Technically, the study has been carried out from two perspectives. The first is proposing robust classification methods with a reject option based on Chow's theoretical works to select ambiguous and outlier samples. The second is proposing robust multi-stage methods for biomedical image classification and detection. To meet the objectives introduced in Chapter 1, work done and presented in this thesis includes three main parts:

(1) This thesis has proposed a novel twin support vector machine with a reject option (RO-TWSVM) to select ambiguous and outlier testing samples.

This part of the thesis aims to propose a novel twin support vector machine with a reject option (RO-TWSVM) to select both ambiguous and outlier testing samples. RO-TWSVM is an extension of the existing support vector machine with a reject option (RO-SVM) [23]. Similar to RO-SVM [23], RO-TWSVM is also formulated based on a cost-sensitive framework which follows the principle of minimization of the expected classification cost. Hence, on the one hand, some advantages of RO-SVM are retained in RO-TWSVM, such as the training for TWSVM hyperplanes and the plot of ROC curve are implemented only once since changes made on cost settings, if any, only influence the slopes of the level curves. Hence, excessive re-calculations for the TWSVM hyperplanes and the ROC curve are not required. On the other hand, RO-TWSVM does not assume that samples being rejected should stay close to the separating hyperplane as RO-SVM does. Since we choose the distance difference in TWSVM to measure the uncertainty of a sample, an outlier sample which is far away from the SVM hyperplane and not rejected by RO-SVM could still be rejected by the proposed RO-TWSVM.

The superiority of the proposed RO-TWSVM is shown by the extensive experiments conducted on the synthetic and real-world benchmark datasets. From the experiments

on two synthetic datasets, it is graphically shown that the proposed RO-TWSVM can produce reject regions of different sizes and locations according to various cost settings. Also, it is shown in the simulations on synthetic dataset 2 called “Cross Planes” that RO-TWSVM generates more reasonable and robust reject regions than RO-SVM due to the flexibility offered by the nonparallel hyperplanes of TWSVM. From the experiments conducted on multiple real-world datasets, it is concluded that the proposed RO-TWSVM significantly outperforms TWSVM-without-RO, and in general, performs better than RO-SVM for datasets which are difficult to be discriminated by one hyperplane.

(2) This thesis has proposed a two-stage method for robust biomedical image classification.

This part of the thesis aims to handle the challenges of biomedical image classification, such as high level noise, image blur, illumination variation and complicated geometric correspondence among various categorical biomedical patterns. To this end, a two-stage method has been proposed for biomedical image classification. At stage 1, we propose a confidence score based classification rule with a reject option for a preliminary decision using the support vector machine (SVM). The testing images going through stage 1 are separated into two groups based on their confidence scores. Those testing images with sufficiently high confidence scores are classified at stage 1 while the others with low confidence scores are rejected and fed to stage 2. At stage 2, the rejected images from stage 1 are first processed by a subspace analysis technique called eigenfeature regularization and extraction (ERE), and then classified by another SVM trained in the transformed subspace learned by ERE. The robustness of the proposed method is achieved due to the three complementary effects: (i) the classification structure is complementary via a reject option method; (ii) the local features adopted (SIFT and SURF) have complementary merits in their robustness to variations in images. (iii) the image representation via different BoW models regarding whether incorporating spatial information is also complementary.

Extensive experiments have been conducted on three benchmark biomedical image datasets. We first compare the proposed two-stage method with each of the individual methods at the separate stages as well as the method which direct fuses the two features. As evidenced by the experimental results, the proposed method effectively integrates the two individual methods of the separate stages. Subsequently, we compare the proposed method with several state-of-the-art local feature based competing methods. It has been shown that the proposed method outperforms the competing methods in terms of multiple classification performance metrics.

(3) This thesis has proposed a two-stage method for robust biomedical image detection.

This part of the thesis aims to investigate robust methods for biomedical image detection. Specifically, we focus on a task of detecting ring-like endosomes in fluorescent microscopy images. To alleviate the limitations of one-stage supervised methods, we propose a two-stage supervised method for the automated detection of ring-like endosomes in fluorescent microscopy images. The proposed method consists of a localization stage cascaded by an identification stage. Given a microscopy image, a number of candidate patches of endosomes are first determined in the localization stage. Subsequently, each of these candidate patches is identified by a pretrained support vector machine from a classification perspective in the identification stage. The proposed method is supervised and therefore it does not require much prior knowledge as unsupervised methods do. The localization stage determines candidate patches and therefore reduces the computational burden of classification. It also alleviates the “mismatch” problem as it aligns each potential endosome at the center of the corresponding candidate patch. Unlike one-stage methods which only exploit global information of training patches, the proposed method applies local information of training patches in the localization stage and global information in the identification stage. Such combination of complementary information significantly improves the detection performance.

The experiments using the real-world microscopy images show that the proposed method can produce significantly better detection results in terms of the lowest log-average miss rate and the highest precision, recall and F_1 score compared with several one-stage supervised methods. This indicates the proposed method produces more accurate detections while lower missing detections of endosomes compared with the competing methods. Compared with unsupervised methods, the proposed method produces fewer false positive detections by exploiting distinct texture features of endosomes.

6.2 Future works

In this section, several possible directions for future works are summarized below.

(1) *Further extensions and applications of RO-TWSVM.*

- **Extend RO-TWSVM for multi-categorical classification.**

The proposed RO-TWSVM in Chapter 3 is for binary classification. It is a promising direction to extend RO-TWSVM for multi-categorical classification. Note that ROC curves play a significant role in our RO-TWSVM formulation but they are usually applied in binary classification. A simple strategy to extend binary classification methods to multi-categorical classification is the one-vs-the-rest strategy [92], i.e., train multiple binary classification models by fixing one specific category and treating the other categories combined as another class. This strategy reduces an M -categorical classification problem to M binary classification problems. Using this strategy, the RO-TWSVM could be applied directly to solve each binary classification problem. It is noted that one-vs-the-rest strategy generates imbalance binary classification problems. It has been shown in [93] that TWSVM is relatively insensitive to imbalance in class sizes as it solves smaller sized quadratic programming problems in order to find non-parallel hyperplanes that pass through the respective class. Compared with

SVM, TWSVM produces more robust classification results in imbalanced datasets, e.g., Electroencephalogram (EEG) signal classification [93]. Therefore, further researches on extending RO-TWSVM to solve multi-categorical classification problems using one-vs-the-rest strategy are promising.

- **Apply RO-TWSVM for multiple instance learning (MIL).**

The second direction about RO-TWSVM is to apply this method to solve the tasks of multiple instance learning (MIL) for image categorization. For an MIL problem, an image is treated as a set containing multiple instances, each occupies a region within the image. The categorization of this image is performed by labeling each of these instances [11]. Ref. [94] has shown that TWSVM performs better and has lower computational complexity in solving MIL tasks compared with several SVM based methods. In [11], support vector machine with a reject option has been applied for solving MIL problems. Motivated by these existing works on TWSVM and RO on solving MIL problems, it is promising to investigate the application of RO-TWSVM for this task.

(2) Modifications on both stages in the proposed detection method for biomedical images

In Chapter 5, we have proposed a two-stage method for biomedical image detection. For both the localization stage and identification stage, we could consider possible ways for further improvement in detection speed and accuracy.

- **Apply unsupervised methods in the localization stage for fast detection.**

In the proposed localization stage in Chapter 5, we locally compared each of the query patches and the testing microscopy image in a supervised fashion to determine candidate patches. This approach is effective but still slow as it needs to determine all the matched pairs of local features extracted from query patches and testing images. In the future, we plan to implement real-time tracking on ring-like

endosomes from a stream of images to study their biological activities. Therefore, increasing detection speed is important.

From the experimental results shown in Section 5.3.4, it is observed that unsupervised methods (e.g., MO-FRST) can achieve fast detection but produce many false positive results. However, having some false positive results is less a problem in the localization stage since we only roughly select candidate targets during the localization stage. Therefore, to speed up the overall detection process, we could consider replacing the supervised method in the localization stage by an unsupervised method, e.g., MO-FRST [70]. It is noted that the identification stage still requires supervised methods to accurately identify the targets from candidate patches.

- **Apply convolution neural networks (CNNs) in the identification stage to improve the detection performance.**

In the identification stage of the detection method in Chapter 5, we apply support vector machine trained on LLC-based bag-of-words histograms to identify endosomes from candidate patches. To further improve the detection performance, we could consider exploiting more advanced image classification model in the identification stage. Recently, convolution neural networks [95] [96] (CNNs) achieve great success in multiple computer vision tasks, e.g., image classification. It has also been shown that CNNs are making major contributions to biomedical image analysis, see, e.g., a review [97]. Therefore, applying advanced CNN models in the identification stage is promising.

Since biomedical image dataset sizes are typically smaller compared to those in nature object computer vision datasets (e.g., hundreds/thousands vs. millions of samples), training a CNN from scratch using biomedical images suffers from overfitting. To fill in the gap between CNN models and small biomedical dataset sizes, transfer learning is applied. Essentially, transfer learning uses pre-trained

CNN networks (typically on natural images) to work on biomedical images. In [98], we investigate classification performance of deep CNNs for microscopy images by exploiting transfer learning and feature concatenation. Motivated by this work, it is promising to design a novel identification of endosome detector using CNNs based on transfer learning to classify endosome patches from non-endosome background patches.

Bibliography

- [1] F. Provost and T. Fawcett, “Robust classification for imprecise environments,” *Machine Learning*, vol. 42, no. 3, pp. 203–231, 2001.
- [2] D. T. Nguyen, W. Li, and P. O. Ogunbona, “Human detection from images and videos: a survey,” *Pattern Recognition*, vol. 51, pp. 148–175, 2016.
- [3] C. K. Chow, “An optimum character recognition system using decision functions,” *IRE Transactions on Electronic Computers*, vol. 7, no. 4, pp. 247–254, 1957.
- [4] C. K. Chow, “On optimum recognition error and reject tradeoff,” *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41–46, 1970.
- [5] T. Pietraszek, “On the use of ROC analysis for the optimization of abstaining classifiers,” *Machine Learning*, vol. 68, no. 2, pp. 137–169, 2007.
- [6] R. El-Yaniv and Y. Wiener, “On the foundations of noise-free selective classification,” *The Journal of Machine Learning Research*, vol. 11, pp. 1605–1641, 2010.
- [7] M. Golfarelli, D. Maio, and D. Malton, “On the error-reject trade-off in biometric verification systems,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 786–796, 1997.
- [8] Z. Zidelmal, A. Amirou, and A. Belouchrani, “Heartbeat classification using support vector machines (SVMs) with an embedded reject option,” *International*

- Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 01, p. 1250001, 2012.
- [9] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. Mesirov, and T. Poggio, "Support vector machine classification of microarray data," 1999.
- [10] G. Fumera, I. Pillai, and F. Roli, "Classification with reject option in text categorisation systems," in *International Conference on Image Analysis and Processing, 2003*, pp. 582–587, IEEE, 2003.
- [11] R. Zhang and D. N. Metaxas, "RO-SVM: Support vector machine with reject option for image categorization.," in *British Machine Vision Conference (BMVC)*, pp. 1209–1218, Citeseer, 2006.
- [12] C. Sommer and D. W. Gerlich, "Machine learning in cell biology—teaching computers to recognize phenotypes," *Journal of Cell Science*, vol. 126, no. 24, pp. 5529–5539, 2013.
- [13] M. V. Boland and R. F. Murphy, "A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells," *Bioinformatics*, vol. 17, no. 12, pp. 1213–1223, 2001.
- [14] R. F. Murphy, M. Velliste, and G. Porreca, "Robust numerical features for description and classification of subcellular location patterns in fluorescence microscope images," *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, vol. 35, no. 3, pp. 311–321, 2003.
- [15] G. R. Johnson, J. Li, A. Shariff, G. K. Rohde, and R. F. Murphy, "Automated learning of subcellular variation among punctate protein patterns and a generative model of their relation to microtubules," *PLOS Computational Biology*, vol. 11, no. 12, p. e1004614, 2015.

- [16] Y. Li, T. D. Majarian, A. W. Naik, G. R. Johnson, and R. F. Murphy, "Point process models for localization and interdependence of punctate cellular structures," *Cytometry Part A*, vol. 89, no. 7, pp. 633–643, 2016.
- [17] T. Peng, G. M. Bonamy, E. Glory-Afshar, D. R. Rines, S. K. Chanda, and R. F. Murphy, "Determining the distribution of probes between different subcellular locations through automated unmixing of subcellular patterns," *Proceedings of the National Academy of Sciences*, vol. 107, no. 7, pp. 2944–2949, 2010.
- [18] L. P. Coelho, T. Peng, and R. F. Murphy, "Quantifying the distribution of probes between subcellular locations using unsupervised pattern unmixing," *Bioinformatics*, vol. 26, no. 12, pp. i7–i12, 2010.
- [19] Y. Hu, E. Osuna-Highley, J. Hua, T. S. Nowicki, R. Stolz, C. McKayle, and R. F. Murphy, "Automated analysis of protein subcellular location in time series images," *Bioinformatics*, vol. 26, no. 13, pp. 1630–1636, 2010.
- [20] A. Shariff, R. F. Murphy, and G. K. Rohde, "A generative model of microtubule distributions, and indirect estimation of its parameters from fluorescence microscopy images," *Cytometry Part A*, vol. 77, no. 5, pp. 457–466, 2010.
- [21] T. Peng and R. F. Murphy, "Image-derived, three-dimensional generative models of cellular organization," *Cytometry Part A*, vol. 79, no. 5, pp. 383–391, 2011.
- [22] G. Danuser, "Computer vision in cell biology," *Cell*, vol. 147, no. 5, pp. 973–978, 2011.
- [23] F. Tortorella, "Reducing the classification cost of support vector classifiers through an ROC-based reject rule," *Pattern Analysis and Applications*, vol. 7, no. 2, pp. 128–143, 2004.

- [24] J. Platt *et al.*, “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advances in Large Margin Classifiers*, vol. 10, no. 3, pp. 61–74, 1999.
- [25] R. Herbei and M. H. Wegkamp, “Classification with reject option,” *Canadian Journal of Statistics*, vol. 34, no. 4, pp. 709–721, 2006.
- [26] P. L. Bartlett and M. H. Wegkamp, “Classification with a reject option using a hinge loss,” *The Journal of Machine Learning Research*, vol. 9, pp. 1823–1840, 2008.
- [27] G. Fumera and F. Roli, “Support vector machines with embedded reject option,” in *Pattern Recognition with Support Vector Machines*, pp. 68–82, Springer, 2002.
- [28] M. Wegkamp and M. Yuan, “Support vector machines with a reject option,” *Bernoulli*, vol. 17, no. 4, pp. 1368–1385, 2011.
- [29] Y. Grandvalet, A. Rakotomamonjy, J. Keshet, and S. Canu, “Support vector machines with a reject option,” in *Advances in Neural Information Processing Systems*, pp. 537–544, 2009.
- [30] B. Hanczar and M. Sebag, “Combination of one-class support vector machines for classification with reject option,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 547–562, Springer, 2014.
- [31] R. Sousa, A. R. da Rocha Neto, J. S. Cardoso, and G. A. Barreto, “Classification with reject option using the self-organizing map,” in *International Conference on Artificial Neural Networks*, pp. 105–112, Springer, 2014.
- [32] L. Fischer, B. Hammer, and H. Wersing, “Efficient rejection strategies for prototype-based classification,” *Neurocomputing*, vol. 169, pp. 334–342, 2015.
- [33] T. Villmann, M. Kaden, A. Bohnsack, J.-M. Villmann, T. Drogies, S. Saralajew, and B. Hammer, “Self-adjusting reject options in prototype based classification,”

- in *Advances in Self-Organizing Maps and Learning Vector Quantization*, pp. 269–279, Springer, 2016.
- [34] F. Tortorella, “A ROC-based reject rule for dichotomizers,” *Pattern Recognition Letters*, vol. 26, no. 2, pp. 167–180, 2005.
- [35] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A training algorithm for optimal margin classifiers,” in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pp. 144–152, ACM, 1992.
- [36] V. Vapnik, *The nature of statistical learning theory*. Springer Science & Business Media, 2013.
- [37] C. J. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
- [38] R. Khemchandani, S. Chandra, *et al.*, “Twin support vector machines for pattern classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 5, pp. 905–910, 2007.
- [39] O. L. Mangasarian and E. W. Wild, “Multisurface proximal support vector machine classification via generalized eigenvalues,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 1, pp. 69–74, 2006.
- [40] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.
- [41] R. Khemchandani and S. Chandra, “Optimal kernel selection in twin support vector machines,” *Optimization Letters*, vol. 3, no. 1, pp. 77–88, 2009.
- [42] L. P. Coelho, J. D. Kangas, A. W. Naik, E. Osuna-Highley, E. Glory-Afshar, M. Fuhrman, R. Simha, P. B. Berget, J. W. Jarvik, and R. F. Murphy, “Determining the subcellular location of new proteins from microscope images using local features,” *Bioinformatics*, vol. 29, no. 18, pp. 2343–2349, 2013.

- [43] D. Liu, S. Wang, D. Huang, G. Deng, F. Zeng, and H. Chen, “Medical image classification using spatial adjacent histogram based on adaptive local binary patterns,” *Computers in Biology and Medicine*, vol. 72, pp. 185–200, 2016.
- [44] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [45] M. Jiang, S. Zhang, H. Li, and D. N. Metaxas, “Computer-aided diagnosis of mammographic masses using scalable image retrieval,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 783–792, 2015.
- [46] F. Mualla, S. Schöll, B. Sommerfeldt, A. Maier, and J. Hornegger, “Automatic cell detection in bright-field microscope images using SIFT, random forests, and hierarchical clustering,” *IEEE Transactions on Medical Imaging*, vol. 32, no. 12, pp. 2274–2286, 2013.
- [47] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (SURF),” *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
- [48] 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*, vol. 24, no. 7, pp. 971–987, 2002.
- [49] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2006*, vol. 2, pp. 2169–2178, IEEE, 2006.
- [50] N. Ali, K. B. Bajwa, R. Sablatnig, S. A. Chatzichristofis, Z. Iqbal, M. Rashid, and H. A. Habib, “A novel image retrieval based on visual words integration of SIFT and SURF,” *PLOS one*, vol. 11, no. 6, p. e0157428, 2016.

- [51] L. Juan and O. Gwun, "A comparison of SIFT, PCA-SIFT and SURF," *International Journal of Image Processing*, vol. 3, no. 4, pp. 143–152, 2009.
- [52] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision (ICCV)*, 2003, pp. 1470–1477, IEEE, 2003.
- [53] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3360–3367, IEEE, 2010.
- [54] X. Shen, Z. Lin, J. Brandt, and Y. Wu, "Spatially-constrained similarity measure for large-scale object retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1229–1241, 2014.
- [55] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.
- [56] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal Component Analysis*, pp. 115–128, Springer, 1986.
- [57] X. Jiang, "Linear subspace learning-based dimensionality reduction," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 16–26, 2011.
- [58] X. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 383–394, 2008.
- [59] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear benchmark data for pattern classification," *Nature Inspired Smart Information Systems (NiSIS)*, 2005, pp. 1–9, 2005.

- [60] P. Foggia, G. Percannella, P. Soda, and M. Vento, "Benchmarking Hep-2 cells classification methods," *IEEE Transactions on Medical Imaging*, vol. 32, no. 10, pp. 1878–1889, 2013.
- [61] J. Newberg and R. F. Murphy, "A framework for the automated analysis of subcellular patterns in human protein atlas images," *Journal of proteome research*, vol. 7, no. 6, pp. 2300–2308, 2008.
- [62] N. A. Hamilton, R. S. Pantelic, K. Hanson, and R. D. Teasdale, "Fast automated cell phenotype image classification," *BMC bioinformatics*, vol. 8, no. 1, p. 1, 2007.
- [63] M. Tahir and A. Khan, "Protein subcellular localization of fluorescence microscopy images: employing new statistical and texton based image features and SVM based ensemble classification," *Information Sciences*, vol. 345, pp. 65–80, 2016.
- [64] B. Zhang and T. D. Pham, "Phenotype recognition with combined features and random subspace classifier ensemble," *BMC Bioinformatics*, vol. 12, no. 1, p. 128, 2011.
- [65] A. Chebira, Y. Barbotin, C. Jackson, T. Merryman, G. Srinivasa, R. F. Murphy, and J. Kovačević, "A multiresolution approach to automated classification of protein subcellular location images," *BMC Bioinformatics*, vol. 8, no. 1, p. 210, 2007.
- [66] I. Smal, M. Loog, W. Niessen, and E. Meijering, "Quantitative comparison of spot detection methods in fluorescence microscopy," *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 282–301, 2010.
- [67] Q. Wu, F. Merchant, and K. Castleman, *Microscope image processing*. Academic press, 2010.
- [68] Y. Kalaidzidis, "Multiple objects tracking in fluorescence microscopy," *Journal of Mathematical Biology*, vol. 58, no. 1-2, p. 57, 2009.

- [69] B. Zhang, M. Fadili, J.-L. Starck, and J.-C. Olivo-Marin, "Multiscale variance-stabilizing transform for mixed-Poisson-Gaussian processes and its applications in bioimaging," in *IEEE International Conference on Image Processing (ICIP)*, 2007, vol. 6, pp. VI-233, IEEE, 2007.
- [70] S. Ram and J. J. Rodríguez, "Size-invariant detection of cell nuclei in microscopy images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 7, pp. 1753–1764, 2016.
- [71] G. Loy and A. Zelinsky, "Fast radial symmetry for detecting points of interest," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 959–973, 2003.
- [72] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [73] S. Jiang, X. Zhou, T. Kirchhausen, and S. T. Wong, "Detection of molecular particles in live cells via machine learning," *Cytometry Part A*, vol. 71, no. 8, pp. 563–575, 2007.
- [74] M. A. Kumar and M. Gopal, "Application of smoothing technique on twin support vector machines," *Pattern Recognition Letters*, vol. 29, no. 13, pp. 1842–1848, 2008.
- [75] Y.-H. Shao, C.-H. Zhang, X.-B. Wang, and N.-Y. Deng, "Improvements on twin support vector machines," *IEEE Transactions on Neural Networks*, vol. 22, no. 6, pp. 962–968, 2011.
- [76] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

- [77] K.-A. Toh and H.-L. Eng, “Between classification-error approximation and weighted least-squares learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 4, pp. 658–669, 2008.
- [78] K.-A. Toh and G.-C. Tan, “Exploiting the relationships among several binary classifiers via data transformation,” *Pattern Recognition*, vol. 47, no. 3, pp. 1509–1522, 2014.
- [79] L. Sun, K.-A. Toh, and Z. Lin, “A center sliding Bayesian binary classifier adopting orthogonal polynomials,” *Pattern Recognition*, vol. 48, no. 6, pp. 2013–2028, 2015.
- [80] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
- [81] D. Lin, Z. Lin, S. Sothiharan, L. Lei, and J. Zhang, “An SVM based scoring evaluation system for fluorescence microscopic image classification,” in *IEEE International Conference on Digital Signal Processing (DSP)*, 2015, pp. 543–547, IEEE, 2015.
- [82] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, Oct 2001.
- [83] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
- [84] I. Mellman, “Endocytosis and molecular sorting,” *Annual Review of Cell and Developmental Biology*, vol. 12, no. 1, pp. 575–625, 1996.
- [85] S. C. Devanaboyina, S. M. Lynch, R. J. Ober, S. Ram, D. Kim, A. Puig-Canto, S. Breen, S. Kasturirangan, S. Fowler, L. Peng, *et al.*, “The effect of PH dependence of antibody-antigen interactions on subcellular trafficking dynamics,” in *MAbs*, vol. 5, pp. 851–859, Taylor & Francis, 2013.

- [86] J. Cao, T. Chen, and J. Fan, "Landmark recognition with compact BoW histogram and ensemble ELM," *Multimedia Tools and Applications*, vol. 75, no. 5, pp. 2839–2857, 2016.
- [87] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005*, vol. 1, pp. 886–893, IEEE, 2005.
- [88] Z. Gan, S. Ram, C. Vaccaro, R. J. Ober, and E. S. Ward, "Analyses of the recycling receptor, FcRn, in live cells reveal novel pathways for lysosomal delivery," *Traffic*, vol. 10, no. 5, pp. 600–614, 2009.
- [89] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [90] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: an evaluation of the state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–761, 2012.
- [91] A. Satpathy, X. Jiang, and H.-L. Eng, "LBP-based edge-texture features for object recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 1953–1964, 2014.
- [92] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, ch. 2, p. 64. John Wiley & Sons, 2012.
- [93] S. Soman, "High performance EEG signal classification using classifiability and the Twin SVM," *Applied Soft Computing*, vol. 30, pp. 305–318, 2015.
- [94] D. Tomar and S. Agarwal, "Twin support vector machine for multiple instance learning based on bag dissimilarities," *Advances in Artificial Intelligence*, vol. 2016, 2016.

- [95] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [96] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.
- [97] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.
- [98] L. D. Nguyen, D. Lin, Z. Lin, and J. Cao, "Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–5, IEEE, 2018.

Author's publications

Journal papers

[1] Dongyun Lin, Lei Sun, Kar-Ann Toh, Jing Bo Zhang, and Zhiping Lin. Twin SVM with a reject option through ROC curve. *Journal of the Franklin Institute*, 355(4):1710 – 1732, 2018.

[2] Dongyun Lin, Lei Sun, Kar-Ann Toh, Jing Bo Zhang, and Zhiping Lin. Biomedical image classification based on a cascade of an SVM with a reject option and subspace analysis. *Computers in Biology and Medicine*, 96:128 – 140, 2018.

[3] Dongyun Lin, Zhiping Lin, Ramraj Velmurugan, Jiuwen Cao, and Raimund J. Ober. A two-stage method for automated detection of ring-like endosomes in fluorescent microscopy images. To be submitted soon.

Conference papers

[1] Dongyun Lin, Zhiping Lin, Ramraj Velmurugan, and Raimund J Ober. Automatic endosomal structure detection and localization in fluorescence microscopic images. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–4. IEEE, 2017.

- [2] Dongyun Lin, Zhiping Lin, Lei Sun, Kar-Ann Toh, and Jiuwen Cao. LLC encoded BoW features and softmax regression for microscopic image classification. In IEEE International Symposium on Circuits and Systems (ISCAS), pages 1–4. IEEE, 2017.
- [3] Dongyun Lin, Zhiping Lin, Shakeela Sothiharan, Lei Lei, and Jing Bo Zhang. An SVM based scoring evaluation system for fluorescence microscopic image classification. In IEEE International Conference on Digital Signal Processing (DSP), pages 543–547. IEEE, 2015.
- [4] Long D. Nguyen, Dongyun Lin, Zhiping Lin, and Jiuwen Cao. Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. In IEEE International Symposium on Circuits and Systems (ISCAS), pages 1–4. IEEE, 2018.
- [5] Xu Zhang, Yunfeng Liang, Dongyun Lin, Zhiping Lin, Steven Tien Guan Thng, Emily Yiping Gan, and Evelyn Yuxin Tay. Reaction-diffusion based level set method with local entropy thresholding for melasma image segmentation. In IEEE International Conference on Control, Automation, Robotics and Vision (ICARCV), pages 1–5. IEEE, 2016.