Nuclear Morphometry based Pattern Recognition in Pathology

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Biomedical Engineering

Chi Liu

B.S., Biomedical Engineering, Central South University

Carnegie Mellon University
Pittsburgh, PA

August, 2017

Keywords: cancer detection, digital pathology, feature extraction, nuclear morphometry, nuclei segmentation, set classification, supervised learning

To my beloved parents

for their unconditional love and support

Acknowledgments

My sincere gratitude goes to my advisor Dr. Gustavo K. Rohde for his patience, invaluable guidance and support of my Ph.D. study. He has taught me how to think problems mathematically, how to write a technical paper and how to be good researcher. He has been my perfect role model as an excellent scientist, a nice advisor and a good friend.

Next, I would like to thank my thesis committee members Dr. Ge Yang, Dr. John Ozolek and Dr. Liron Pantanowitz for participating in this work, for their support and for insightful comments and suggestions.

I would also like to express my gratitude to our collaborators Rajendra Singh, Oleksandr Yergiyev, Matthew G. Hanna, Zoltan Oltvai, Huazhang Guo and Sanja Dacic who provided valuable data for my projects and spent time teaching me a lot of basic knowledge in pathology. I am grateful for the opportunities that I had to work with many talented colleagues over the past years, including Dr. Liam Cattell, Dr. Akif Burak Tosun, Dr. Saurav Basu, Dr. Soheil Kolouri, Dr. Serim Park, Dr. Shinjini Kundu, Dr. Fei Shang, Dr. Liang Ge, Dr. Shuchang Xu, Dr. Yue Huang, Kan Jia, Ligong Han, Yang Zou and Dr. Siheng Chen. Thank you for your encouragements and useful discussions. Most of all, I would like to thank my parents and sister for their unconditional love.

Last but not least, I would like to thank my funding sources. This work was made possible by grants NIH R01GM090033 and R21CA188938.

Abstract

Given the strong association between aberrant nuclear morphology and tumor progression, changes in nuclear structure have remained the gold standard for cancer diagnosis for over 150 years. Recently, the rapid development of imaging hardware and computation power creates the opportunity for automated computer-aided diagnosis (CAD). Developing a robust and reliable pattern recognition pipeline is a pressing need to mine and analyze tons of nuclei data being captured.

Among the rich studies on pattern recognition problems in pathology, automated nuclei detection, segmentation and cancer detection are the recurring tasks due to the importance and challenges of nuclei analysis. In this thesis, we propose and investigate the state-of-art methods in the CAD modules for maximizing the overall amount of information from images for decision making. We focus on nuclei segmentation and patient cancer detection in the nuclei image analysis pipeline.

As the first step in nuclei analysis, we develop an unsupervised nuclei detection and segmentation approach for pathology images. Different from many supervised segmentation methods whose performances rely on the quality and quantity of training samples, the proposed method is able to automatically search for the nucleus contour by solving the shortest path problem with little user effort. We consider the cancer detection task as a set classification problem and propose a highly discriminative predictive model in the sense that it not only optimizes the classifier decision boundary but also transfers discriminative information to set representation learning. The innovation of the model is the integration of set representation learning and classifier training into one objective function for boosting the cancer detection performance. Experimental results showed that the new model provides significant improvements compared with state-of-art methods in the diagnostic challenges. In addition, we showed that the predictive model enables visual interpretation of dis-

criminative nuclear characteristics representing the whole nuclei set.

We believe the proposed model is quite general and provide experimental validations in several extended pattern recognition problems.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and motivation

Nuclear architecture provides a framework for regulating numerous functions in the nucleus, which is recognized as the site for storage and organization of the genetic material, DNA synthesis, DNA transcription, transcriptional regulation, and RNA processing in eukaryotes [18]. The change of nuclear architecture can be spurred by aberrations in the genetic code and the transcription of different messenger RNAs related to biological processes of cancer [72]. Today, cancer is the second leading cause of death in the US and a major public health problem worldwide, accounting for 8.8 million deaths in 2015 [31]. Even though the biology of cancer lacks completely understanding, nuclear architecture in cancer cells has been found to show characteristic differences compared with normal cells. Given the strong association between aberrant nuclear morphology and tumor progression, changes in nuclear structure have remained the gold standard for cancer diagnosis for over 150 years [13]. Many tumors have characteristic nuclei alterations which can be manually analyzed by pathologists in therapeutic decision making. Such morphological alterations including changes in nuclear size, shape, appearances of nucleoli and chromatin arrangement, provide an important diagnostic feature [104] (See Figure 1.1).

In clinical diagnosis, histology and cytology are two kinds of imageries seeking to examine

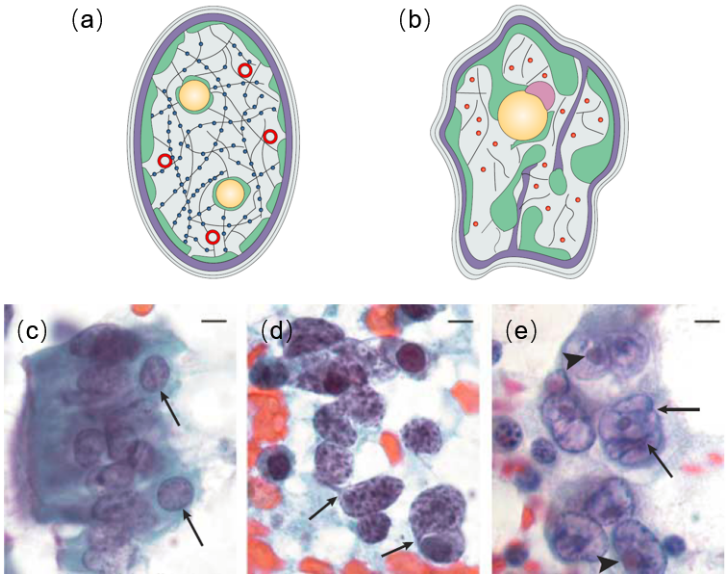

Figure 1.1: Exemplary nuclear structure differences in normal and cancer cells. (a) Normal nucleus bounded by nuclear lamina, a proteinaceous layer made of the lamins and associated proteins; (b) Cancer nucleus with changes in shape, chromatin aggregation, nucleoli and so on; (c) Normal bronchial cells; (d) Small-cell lung carcinoma; (e) Large-cell lung carcinoma. This figure is taken from [104].

the structure of tissues (histology) and cells (cytology and histology) at the microscopic level. After a sequence of technical procedures for preparation, the characterization of nuclear morphology can be visually interpreted under light microscopy with cells stained with reagents (*e.g.* Hematoxylin and Eosin, Feulgen, Diff-Quik) [40]. As in the past decades, manual analysis of nuclear morphology in pathology images still remains the primary approach to determine the presence or absence of cancer for patients, which heavily depends on the personal expertise of pathologists. However, the sheer volume and complexity of nuclear apperances displayed in pathological images make visual interpretation a daunting task for the human and represent laborious work for pathologists. It is extremely difficult to memorize and analyze the distribution of nuclear morphology for thousands of cells located at distinct sites in the slides to gain insights into disease progress. Further more, such manual analysis of sample slides is subjective and often lead to considerable variability. Diagnostic discrepancies happen even for relatively common diseases among pathologists due to the inter-observer variability (differences between pathologists when interpreting the same slides) and intra-observer variability (differences in how

one individual interprets slides at different times or the same lesion represented multiple times) [41]. This may bring ramifications to patients on cancer diagnosis, prognosis prediction and personalized medicine.

Thanks to the adoption of high-resolution imaging systems, glass slides can be converted into digital pathology images with diagnostic quality, which creates the opportunity for automated computer-aided diagnosis. Computerized image analysis can provide quantitative assessments of anatomic entities in pathology images and has received a lot of attention in digital pathology with growing applications related to nuclear morphometry. With the increasing imaging resolution and computation power in hardware, there is a pressing need to develop robust and reliable pattern recognition pipeline to mine and analyze tons of nuclei data being captured. The endeavour in the area of pattern recognition will in turn facilitate the utilization of digital imaging systems in pathologists' workflow with the ultimate goal of minimizing human intervention, reducing turnaround time, and providing traceable clinical information.

Among the rich studies on pattern recognition problems in pathology, automated nuclei detection, segmentation and cancer detection are the recurring tasks due to the importance and challenges of nuclei analysis [40]. In this thesis, a computer-aided diagnosis (CAD) pipeline is developed for detecting and visualizing nuclear morphological differences from pathological images.

We propose and investigate the state-of-art methods in the CAD modules for potentially maximizing the overall amount of information extracted from nuclei images for decision making. Specifically, given the input pathological images, nuclei are automatically detected and isolated from the background structures in an unsupervised manner. After that, nuclear morphology is characterized by a feature vector (*e.g.* using transport-based morphometry, a geometric approach by considering the distribution of pixel intensity over the image coordinates). Finally, patient-level prediction modeling and exploratory analysis are applied to improve cancer detection performance. Experimental validations confirmed that the proposed nuclei image analysis pipeline is potentially practical in building accurate, automated and interpretable CAD systems

for cancer detection tasks. We believe that the proposed predictive model is quite general and can thus be applied to many pattern recognition problems in the biomedical domain facilitating differentiation of characteristic patterns associated with the class.

1.2 Previous work on nuclei image analysis

The earliest work of computer-aided diagnosis can be dated back to the use of digital mammograph in 1990s [34]. Today, the automated recognition of pathological patterns enables fast, quantitative and reproducible characterization of nuclear morphology and has achieved a certain degree of success not only in clinical usage (e.g. cancer detection [83], staging [23], prognosis prediction [96]) but also in cancer research (e.g. drug discovery [100]).

Figure 1.2: Typical nuclei image analysis pipeline.

As shown in Figure 1.2, the typical pattern recognition pipeline for nuclei image analysis consists of three main modules: nuclei segmentation (including image normalization, color deconvolution, nuclei detection and delineation), nuclear morphometry quantification and data analysis (*e.g.* predictive model learning, visualization, *etc*.). The importance of quantitative and automated pipeline has led to several commercial or open-source software tools such as CellProfiler (Broad Institute), GENIE (Aperio, Vista, California, USA), ImageJ (National Institutes of Health), Definiens-Tissue Studio (Definiens, Munich, Germany), HALO (Indica Labs, New Mexico, USA), mitoSEK (Inspirata, Florida, USA), AQUA Analysis (HistoRx, Connecticut, USA), and Visiopharm (Hoersholm, Denmark) [50].

In addition to the mentioned software tools, in sections below we provide a brief review of

methods for nuclei segmentation and predictive modeling for cancer detection in the literature. Refer to for [40], [34], [50], [46] for a review of pre-processing techniques such as image normalization and denosing. Here we only highlight existing state-of-art solutions in key components of the nuclei-based pattern recognition pipeline.

1.2.1 Review on nuclei segmentation in pathology images

Detecting and segmenting nuclei correctly with minimum human effort is a critical prerequisite in CAD system and is important for subsequent nuclei analysis in the pipeline. For example, Chanho *et al*. [44] showed that improved segmentation accuracy led to better classification performance for thyroid follicular lesions using the unique classifier.

Nuclei detection plays a critical role in the overall segmentation procedure, which requires a point per nucleus and close to nucleus center, referred to as seed. Many approaches have been described in the literature to locate nuclei in 2D microscopy images. Distance transform, morphological operation, H-minimum/maximum transform, LoG filtering, Hough transform, radial symmetry transform and machine learning-based methods are the major methods in the literature [90]. For better nuclei detection performance, the variants or combination of these algorithms are designed for specific tasks. The combination of finding peaks in the Euclidean distance map and watershed [72], though often resulting in over-seeding, can be applied to locate seeds. The circular shaped nuclei can be effectively located using Hough transformation methods at the cost of expensive computation [17]. H-maxima/minima transform is a powerful approach to detect nuclei by suppressing spurious local intensity maxima/minima and have been applied in nuclei detection in Pap smear images [68], FISH images [71] and IHC-stained breast cancer images [61]. However, it often leads to overseeding due to its sensitivity to image textures. In image analysis, LoG filter is one of the popular methods for blob detection. The multi-scale Laplacian-of-Gaussian filtering constrained by the distance-map-based adaptive scale selection can be used to detect cell nuclei [2]. Qi *et al.*[69] proposed a method based on single-path voting followed by mean-shift clustering to find seeds for touching and overlapping nuclei. Nowadays, machine

learning-based methods have been proposed to deal with the rich variety of nuclei appearances in pathology images. Support vector machine (SVM), random forest, deep neural networks are frequently used to detect nuclei in prostate [47], bladder [60], breast cancer images [14], *etc*. It is worth mentioning that recently deep learning methods (especially convolutional neural network [14], sparse autoencoder [93]) have attracted a great deal of attention in nuclei detection for its nature of automated feature extraction. Other nuclei detection methods include clustering (mean-shift clustering [16], Fuzzy C-means clustering [4]), template matching [12], and dictionary learning [78].

Nuclei segmentation is to extract individual nuclei from the surrounding structures by delineating their real boundaries. Nuclei segmentation has been extensively studied in the past decades and new segmentation techniques will continue to be proposed for applications including cancer detection and grading. Thresholding, morphological operation, region-based methods, watershed, deformable models, clustering, graph-based methods and supervised classification are the cornerstones of the segmentation methods proposed in the literature.

Thresholding is the most intuitive segmentation method which needs one global threshold or multiple regional thresholds to convert the gray-scale image into the binary image. The performance of thresholding highly depends on the choice of the threshold and the distribution differences between nuclei intensity and background [90]. Otsu's method [65] aims to automatically select the optimal threshold by minimizing the intra-class variance. To deal with nonuniform illumination, an image can be divided into subregions where local adaptive thresholds are computed for binarization. Though simple and fast, thresholding suffers from including little object knowledge and lacking robustness to size, shape as well as texture variations [40].

Morphological operation including basic operations such as erosion, dilation, opening and closing is often combined with other methods for nuclei segmentation. In [33], morphological operations and thresholding were combined to segment nuclei in neuroblastoma images. A multiscale decomposition method was proposed based on mathematical morphology operation in [75] for cell segmentation which is invariant with cell cluster size.

Region-based methods are simple and fast to segment an image into regions directly. Region growing [20], region splitting [64] and region merging [19] are three common region-based segmentation approaches. Region growing usually begins with seed points and finds the region of interests by examining surrounding pixels based on predefined similarity criterion. The idea of region merging is to merge small regions to neighboring larger regions with similar characteristics to avoid over segmentation. Region splitting is the opposite of region merging and starts with the image as the single region. It recursively divides the image into subsidiary regions until the condition of homogeneity is satisfied.

Watershed algorithm is a commonly used nuclei segmentation approach which requires predetected nuclei seeds. The basic idea is to view the image as a topographical relief and the pixel intensity as the elevation. The landscape is gradually flooded with water from regional minima and dams are built to prevent water in different basins from merging [90]. The dam boundaries in the landscape are watershed lined and used to separate image regions. The disadvantage of watershed is that direct use of watershed algorithm is likely to produce over segmentation results due to intensity variations of nuclei and background. Therefore, marker controlled watershed algorithm is often used for segmentation where a marker is a connected component corresponding to an object to be segmented [57].

Deformable models have been widely used in biomedical image segmentation [77],[97], [103] with satisfying performances. Deformable models usually begin with a initial position (manually initialized in many cases) and then gradually evolve toward to the object boundary under the control of internal force and external force. The internal force is to constrain smoothness of the contour, while the external force is to drive the contour to the boundary of the interested object. The deformable models can be classified into two categories: geodesic models and parametric models [90]. In geodesic models, the contour is implicitly represented as the zero level set of a high-dimensional manifold with the benefit of following topology changes naturally. To segment hundreds of nuclei simultaneously, the well-known Chan-Vese (CV) [10] model is often used with one initialization per nucleus. In the parametric model, the contour is explicitly

represented as in parametric form while deforming. Active contour model, or 'snakes' [95] is a classic model which has lower computation cost and can thus be solved very fast compared with geodesic models. However, parametric models can't deal with topology changes (*e.g.* splitting, merging) during the process of deformation.

Clustering methods are a set of algorithms that group a collection of instances into subclusters such that in a certain space instances within the same subcluster are closer than those from distinct subclusters. One fundamental problem shared in clustering methods is the selection of similarity metric for instance-wise distance. Metrics like Euclidean distance, city block distance, Minkowski distance, correlation and 0-1 error are commonly used in cluster analysis [90]. In the application of nuclei image segmentation [40], many clustering approaches are designed based on three basic clustering methods: *k*-means, fuzzy c-means and expectation maximization (EM) algorithm. The *k*-means clustering associates each of the instances with only one subcluster through hard assignment, while fuzzy c-means allows each instance to belong to more than one subclusters using the membership degree. The EM algorithm for Gaussian mixture model is one of the most widely used method for image segmentation by assuming that in a feature space pixels are generated from a mixture of a finite number of Gaussian distributions with unknown parameters [26].

Graph-based methods perform image segmentation by modeling an image as a weighted undirected graph where pixels/superpixels are graph nodes and edge weights are the similarity between pair-wise nodes. The graph can be partitioned into multiple sets for image segmentation based on a certain criterion. Max-flow/Min-cut, normalized cut, conditional random field (CRF) and random walk are typical graph-based approaches. Kofahi *et al.* [2] utilized graph-cut-based binarization to extract the foreground, and then a second graph-cut-based algorithm to refine the initial contours obtained by constrained multi-scale LoG filter, which was shown to perform well in pathology images with dense nuclei. CRF usually formulates nuclei image segmentation as a classification problem, where latent labels of graph nodes are infered based on the observations [89].

Supervised classification methods are increasingly proposed recently and seek to learn machine learning models from labeled exemplars with domain knowledge to deal with the complexity of nuclei image data. Based on the types of input samples in classification, classification-based segmentation can be classified into pixel-wise classification and superpixel-wise classification. In pixel-wise classification, models learn from pixel properties and assign labels (e.g. foreground, background) to each pixel for segmentation. However, pixel-classification is unable to handle touching objects and requires post-processing operations to separate pixel clusters into individual nuclei [49]. On the contrary, superpixel-wise classification first partitions the image into a collection of small candidate regions based on some properties and learn high-level representations in a certain feature space [6]. It has lower computation cost compared to pixel-wise classification, but its performance highly depends on the quality of generated superpixels [90]. In supervised classification, one classifier or a set of aggregated classifiers are trained for label prediction.

1.2.2 Review on patient-level predictive models

In clinical diagnosis, pathologists rely on microscopic examination of a set of nuclei within the tissue sample for analysis. Thus, in most situations, a diagnostic label is only available for the tissue sample rather than individual nuclei. A predictive model is required to learn from sets of nuclei without nuclei-level annotations and predict the diagnostic label for a new set of nuclei, referred as set classification problem. Beyond cancer diagnosis, set classification problem is also ubiquitous in prognosis prediction, where the model needs to predict the patients survival outcome by taking account of a set of quantified nuclei [96]. Different from conventional image classification where training and testing samples are labeled single-shot images, in the set classification scenario, training and testing samples are sets, each of which consists of various numbers of unlabeled nucleus images. The set classification problem is challenging and cant directly solved by supervised machine learning approaches. Existing solutions to nuclei set classification in the literature are described as follows.

Though often implicitly, many predictive models solve the set classification problem with single image classification by making specific assumptions regarding the relationship between the set label and the distribution of its belonging instance labels. Many studies assume that at least half of the instances in a set represent the set label and thus apply the majority voting strategy in set prediction. Predictive models using majority voting have been described for diagnosing a wide variety of cancers including lung cancer [98], cervical cancer [67] and breast cancer [27], to name a few. In [5], a threshold-based voting strategy was adopted for hepatocellular carcinoma tumor grading. However, the voting threshold for a set being categorized into a certain class needs to be pre-defined based on domain knowledge for the best performance. In the multiple instance learning (MIL) framework [3], one set is considered positive when there is at least one positive instance within it, otherwise the set is considered negative. The standard MIL has attracted a wide range of interests and has been applied successfully in the medical diagnosis domain [94], [70].

Set classification considers the set information as a whole and learns the predictive model at the set level. We note that the idea of classifying nuclei sets instead of individual nuclei is not new. In general, such approaches can be divided into two categories according to whether the set-level information is extracted explicitly. In the first category, the global set information is extracted implicitly by measuring the distance/similarity between two sets. Together with set labels, distance-based classifiers (e.g. K- nearest neighbor, support vector machine) can be trained to predict the unknown set label. Besides straightforward definitions of set-set distance (e.g. Hausdorff distance [85], earth mover distance [39]), in the pattern recognition field, other forms of distance between sets have been proposed to set classification problems. In [9], each set is represented as a convex geometric region spanned by its instances in the feature space and set distance is defined as geometric distances between convex models. In [30], set-wise distance is defined as the sum of local kernels for pairwise instances for two sets, where the type of kernel can be polynomial, radial basis and so on. In [84], the matching kernel was proposed for object recognition based on the idea of maximizing the similarity between two sets. In [101], each set

is mapped to an undirected graph with its instances being the graph nodes and the set distance is then measured by a pre-defined graph similarity function.

In the second category, methods are based on the idea of representing the entire set explicitly with a feature vector [3]. As a result, each set is mapped to an embedded feature space by a mapping function, where standard classifiers can be trained for prediction. In nuclei-based cancer detection, one popular method is to aggregate multiple statistics about feature attributes of nuclei within a set. Statistics [34],[24] such as mean, maximum, minimum, standard deviation, median, are frequently used to summarize the characteristics of the nuclei set. Another type of methods, which we call prototype-based approaches, seek to provide the set embedding by quantifying the presence or similarity with respect to pre-defined prototypes in a particular set. The prototypes can be defined either in instance space or in set space. Bag-of-words method [35], [66] is a typical example, which learns a number of representative instances (prototypes) in the training set and then provides a histogram about the composition of any set in terms of each prototype. Unsupervised clustering, such as k-means, is usually adopted in feature space to generate a collection of cluster centers as dictionary.

1.3 Contributions

Even though a plethora of studies have been published in the field, the automation and accuracy of cancer prediction still needs to be improved. In this thesis, we focus on proposing novel algorithms for two key components in the nuclei image analysis pipeline: nuclei segmentation and cancer prediction for patients. Our specific contributions in this thesis are:

· **Contribution 1: Developing an unsupervised nuclei detection and segmentation approach for pathology images.**

Segmentation is an essential stage in systems for quantitative analysis of nuclei extracted from pathology images. In the literature, the effectiveness of many supervised nuclei segmentation methods largely depends on the quality and quantity of training samples. One example is the

recently proposed convolutional neural network (CNN), which usually requires a large number of labeled instances for training. In addition, many trained CNN models lack generality and are rarely validated on segmenting nuclei from unseen datasets. In this thesis, we propose an unsupervised nuclei segmentation approach based on the observation that strong edge and intensity consistencies exist among neighbor pixels along the nucleus contour. We solve the nuclei segmentation problem by finding the shortest path between two nodes in an undirected graph. The method has been validated on several types of nuclei datasets with different stainings.

· **Contribution 2: Developing a discriminative predictive model for cancer detection**

In nuclei based cancer detection, the ultimate goal is to predict whether a given patient has cancer or not based on the morphology information from a set of extracted nuclei. Many predictive models train a classifier at nuclei level by making certain assumptions regarding the relationship between patient label and the distribution of its belonging nuclei labels. For any test patient, the diagnostic label is assigned by a voting strategy based on nuclei level predictions. However, it is reasonable to expect that cancerous tissues may contain a portion of cells displaying normal phenotypes in addition to cells exhibiting abnormal phenotypes. In this thesis, we consider the nuclei belonging to one patient as a whole and formulate cancer detection as a set classification problem. The method directly builds a predictive model at patient level avoiding the need to make any assumption. The cancer detection performance of the proposed method has been validated on liver cancer, thyroid cancer and melanoma with multiple nuclear quantification approaches.

· **Contribution 3: Providing experimental validations for the predictive model in general pattern recognition problems**

Predicting the class label for a set of instances is an important and ubiquitous problem with many applications where data is captured by various types of sensors. We believe the proposed predictive model is quite general and thus can be applied in many pattern recognition tasks beyond nuclei-based cancer detection in pathology. Similar to cancer detection with nuclei images, we consider these extended problems under the set classification framework, where a set consists

of various number of instances with only set level labels available. We demonstrated the effectiveness of the predictive model in tasks of mass classification in mammograms, classification of flow cytometry data and natural scene classification. In addition, the proposed model alleviates the effort to access instance level annotations in many situations.

1.4 Outline

The rest of the thesis is organized as follows:

The second chapter of the thesis introduces the proposed unsupervised method for nuclei detection and segmentation in 2D pathology images. We compared the method with both supervised and unsupervised segmentation approaches on datasets with different stainings. Qualitative and quantitative analysis showed that the method is automatic and accurate for segmenting nuclei from pathology images with noisy background and has the potential to be used in clinic settings.

In the third chapter of the thesis, we describe a novel approach for set classification in nuclei-based cancer detection. We demonstrated that the proposed model outperforms several state-of-art approaches using three types of cancer datastes. The validations with different nuclear features suggested that SetSVM is likely to provide superior performances independent of nuclear quantification approaches. In addition, we explored the ability of the predictive model to visually interpret discriminative nuclear characteristics representing the patients.

The fourth chapter of the thesis aims to extend the proposed predictive model to several general pattern recognition problems including mass classification in mammograms, flow cytometry data classification and natural scene classification. The experimental validations confirmed that the model enables better separation between different data classes regardless of the types of data and instance measurements.

Finally, Chapter five concludes the thesis and list future work and directions in the area.

Chapter 2

Detecting and Segmenting Nuclei in Two-Dimensional Pathology Images

2.1 Introduction

As mentioned in section 1.2.1, segmentation is an essential stage in systems for quantitative analysis of nuclei extracted from microscopy images. Given the wide variety of nuclei appearances in different organs and staining procedures, a plethora of methods have been described in the literature to improve the segmentation accuracy and robustness. Nuclei segmentation can be classified into unsupervised and supervised approaches based on whether labeled data is required for model training. Supervised methods are proposed to handle the large variability of nuclei appearances in the image and require manually labeled samples to produce an inferred function for mapping new samples. For example, deep learning models learn the highly non-linear mapping function in a supervised manner and have been applied to nuclei detection [93], [91] and segmentation [42]. However, deep learning models often require a large amount of labeled data for model training. In [91], more than 1.5 million labeled nuclei patches were used to train the convolutional neural network (CNN) for nuclei detection and shape initialization. Moreover, deep learning models may need retraining when applied to unseen pathology images with very different nuclei appear-

ances. In [93], a stacked sparse autoencoder model was designed for detecting nuclei in breast cancer images but without validation on other types of nuclei images.

In this chapter, we describe an unsupervised nuclei segmentation method without the requirement of manual annotations, which we call MESPS (multi-scale edge selection in polar space). Specifically, a filter bank consisting of rings with various sizes is first constructed. Nuclei seeds are located by finding the local maximums in the response map generated in normalized cross correlation. In the segmentation step, nuclei contours are iteratively refined by selecting the correct edges in polar space at different smoothing levels. The produced final contour would attach tightly to the actual nucleus border. Figure 2.1 shows the overview of the proposed method. We believe the accurate nature of the segmentation procedure, the simplicity of use and computational efficiency are key advantages of our method as will be demonstrated. The validation study was conducted over two nuclei datasets with ground truth, including 25 Hematoxylin and Eosin (H&E) stained liver histopathology images, and 35 Papanicolaou stained thyroid images. The nuclei detection accuracy was measured by miss rate and the segmentation accuracy was evaluated by two types of error metrics. Overall, the nuclei detection efficiency of the proposed method is similar to the supervised template matching method. In comparison to four state-of-art segmentation methods, the proposed method performed the best with average segmentation error of 10.34% and 0.33 measured by AER and NSD ($10\times$) respectively. Quantitative analysis showed that the method is automatic and accurate when segmenting nuclei from microscopy images with noisy background and has the potential to be used in clinic settings.

2.2 Methods

2.2.1 Nuclei detection

The basic idea of nuclei detection is to find the evidence of presence or absence for a nucleus contained in local image regions. To that end, we construct a filter bank composed of rings with different sizes modeled by the function: $r^2 \leq x^2 + y^2 \leq (r + \zeta)^2$ where r is the radius and ζ

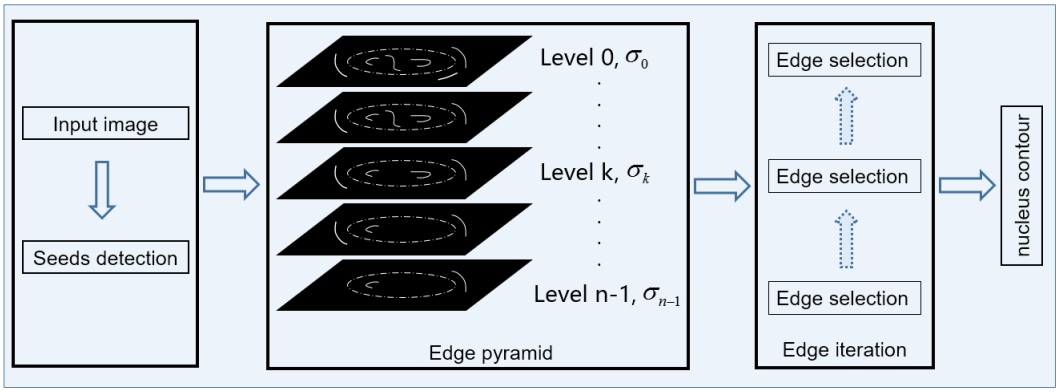

Figure 2.1: Overview of the nuclei detection and segmentation procedure. The nuclei seeds are firstly detected using a set of filters with different sizes. An edge pyramid is then constructed, where edge maps are generated using a set of smooth parameters. Edge selection is performed at each level and the nucleus contour evolves across the edge pyramid to delineate the spatial content of nuclei.

is the thickness. Given a certain data set, prior information such as ζ, the size of the smallest and largest nuclei can be reasonably estimated, and the size range of the filters can be defined according to image resolution. Different functions can change and model the shape of filters to adapt to various nuclei appearances in the datasets. In our experiment, the sampled locations $\bar{x} = [x_i, y_i]$ can be obtained from a set of centered coordinates $[x_1, ..., x_{2r+1}]$, $-r-\zeta \leq x_i \leq r+\zeta$.

The image patch with filter size r denoted as $f_r(\bar{x})$, is convolved with a Gaussian function, which is meant to be an approximation of point spread function (PSF). Given an image $I(\bar{x})$ the likelihood of a pixel being the center of an underlying nucleus is proportional to the following:

$$L(\bar{x}^*) = I(\bar{x}^*) \max_r \{I \circ f_r(\bar{x}^*)\} = I(\bar{x}^*) \max_r \{\frac{\sum_{\bar{x}} I(\bar{x}) f_r(\bar{x} - \bar{x}^*)}{\bar{I}(\bar{x}^*) \bar{f}_r(\bar{x})}\} \tag{2.1}$$

where $\circ$ denotes the normalized cross correlation (NCC) between the filter $f_r(\bar{x})$ and the image I. $\bar{f}_r(\bar{x}) = (\sum_{\bar{x}} (f_r(\bar{x}))^2)^{\frac{1}{2}}$ and $\bar{I}(\bar{x}^*) = (\sum_{\bar{x} \in \Omega} (I(\bar{x}))^2)^{\frac{1}{2}}$, with Ω being the neighborhood of pixel $\bar{x}^*$ with the same size as the filter $f_r(\bar{x})$.

The maximization procedure mentioned above is performed pixel by pixel searching for the filter $f_r(\bar{x})$ within the filter bank which best matches the appearance of the potential nucleus at location $\bar{x}^*$. Pixels with ring shaped surrounding neighborhoods and with similar radius as that of a filter will have strong responses and are likely to be nuclei centers. On the contrary, irrelevant

tissue structures or noisy background tend to have weak responses. Thus, the method is not only able to locate potential nuclei but also yield size estimation for the nucleus by searching for the best matched filters.

Figure 2.2: Simulation for detecting both circular and elliptical nuclei with ring shaped filters. (a) Constructed filter bank with filters at different sizes (magnified for viewing purpose). (b) Simulated microscopy image with circular nuclei. (c) Response map for (b). (d) Simulated microscopy image with elliptical nuclei. (e) Response map for (d).

Since most nuclei in the slides take the shape of an ellipse, elliptical filters would theoretically generate stronger responses compared with ring shaped filters. However, more parameters (*e.g.* length of major and minor axis, rotation angle) are required to control the shape of an ellipse, leading to a larger parameter searching space when performing the NCC operation and generating the response map for nuclei detection. Here we use the ring shape filters instead. As shown from the simulation experiment in Figure 2.2, ring shaped filters are able to generate the strong responses when applied to detect both circular and elliptical nuclei in noisy background.

To locate the nuclei seeds with the response map, the standard k-means clustering method is applied to classify the image pixels into three classes based on their corresponding intensities: 1) background; 2) weak responses from non-nuclei structures; 3) strong responses from potential nuclei. Using connected component analysis, the location of nucleus seed can be obtained by computing the mass center of each isolated pixel cluster classified as strong responses. Figure 2.3 shows the nuclei detection procedure applied to the real liver histopathology images.

In practice, post-processing operations such as thresholding the area of isolated pixels clusters are required to filter out the false positive nuclei seeds.

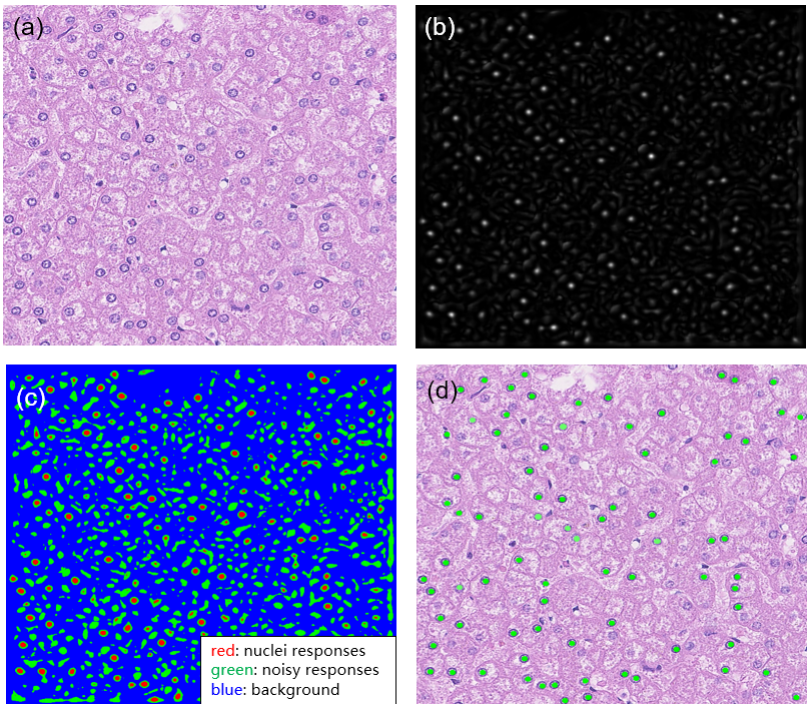

Figure 2.3: Nuclei detection on real nuclei image using the proposed method. Here we separated the Hematoxylin channel from the original RGB color space by color deconvolution [74]. (a) Original liver histopathology image. (b) Response map after normalized cross correlation. (c) k-means clustering results (in colors). (e) Detected nuclei seeds marked as green dots.

2.2.2 Nuclei segmentation

With detected nuclei seeds, it is desired that the subsequent segmentation algorithm delineate the nuclei contours efficiently and accurately with minimal manual intervention. Our goal is to segment nuclei correctly in the complex background (caused by the large variety of nuclei shapes, chromatin textures, staining procedure as well as tissue heterogeneity). Edge detectors can preserve important structural properties of the image and produce boundaries precisely at locations with relatively large gradients, *e.g.* nuclei borders and apparent noisy background structures (Figure 2.3(a)). In order to obtain the initial segmentation, a blurred version of the nuclei image is required, which describes the nuclei outlines and excludes noisy details challenging the delineation of nuclei contours. The multi-scale strategy enables the nucleus contour to refine iteratively from the initial segmentation by changing the blur parameter σ smoothly. The idea of proposed method is to discriminate the nucleus border pixels from remaining 'garbage pixels'

and locate the contour gradually in an iterative manner.

Specifically, the input image is first convolved with the 2D Gaussian function $g(x, y; \sigma_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2+y^2}{2\sigma_i^2}}$ with zero mean for denosing. Parameter $\sigma = [\sigma_0, ..., \sigma_{n-1}]$ indicates the smooth level in the multi-scale strategy with σ_0 and σ_{n-1} being the minimum and the maximum of σ respectively. An edge pyramid is constructed consisting of a set of edge maps generated by the edge detector (*e.g.* Canny detector), where the top level and the bottom level correspond to σ_0 and σ_{n-1} respectively. We aim to select the correct edges in polar space at each smooth level and then take it as guidance for edge selection in the next higher level. The algorithm refines the contour iteratively starting from the bottom level and produces the final contour when edge selection is performed at the top level of the edge pyramid.

Edge selection

The algorithm selects correct edge pixels in the edge map starting from the largest scale σ_{n-1}, where artifacts are the least prevalent. The size of the image patch for each nucleus can be determined adaptively according to the size estimation from nuclei detection step. In the nucleus edge map, edge segments can be classified into three categories: correct edges forming the nucleus contour, edges inside the nucleus, and edges outside the nucleus.

One intuitive and prominent feature to discriminate these three kinds of edge segments is that correct edge pixels on the nucleus contour often have smoother distance changes away from the seed in comparison to the drastic distance fluctuations of pixels on noisy edges inside or outside the nucleus. In addition, considering the intensity, edge pixels along the nucleus border have relatively consistent intensity compared with that of pixels on discontinuous edge segments. Based on these observations regarding edge pixels' locations and intensity, the solution of delineating nucleus contour becomes finding the path with the minimal transportation cost (non-zero) starting and ending at any chosen point on the nucleus border based on both distance and intensity metrics.

The polar coordinate system provides a natural space to search for the optimal path connect-

ing the start point and the end point for each nucleus. Given the edge map E_{σ_i} detected at the *i*th level, edge pixel $\bar{x} = [x, y]$ in Cartesian coordinate can be transformed into polar space by: $r = \sqrt{(x - x^*)^2 + (y - y^*)^2}, \theta = arctan\frac{y-y^*}{x-x^*}, \theta \in [0, 2\pi]$, where $\bar{x}^* = [x^*, y^*]$ is the coordinate of nucleus seed in the image patch. Figure 2.4(b) shows the transformed pixels in polar space originally from the edge map in sub window of Figure 2.4(a).

In the polar coordinate system, the transportation cost between any two neighbor edge pixels $p_m = [r_m, \theta], p_n = [r_n, \theta + \Delta_\theta]$ is defined as follows:

$$\begin{aligned} c(\theta, p_m, p_n) = \quad & \alpha\frac{|r_m - r_n|}{R_{max}} + \quad b\frac{|v_m - v_n|}{V_{max}} \\ s.t. \quad \alpha \geq 0; \quad & b \geq 0; \quad \alpha + b = 1 \end{aligned} \tag{2.2}$$

where v_m and v_n denote pixel intensities for p_m and p_n; $\Delta\theta$ is infinitesimal; α and b are the weights for the distance term and intensity term respectively; R_{max} and V_{max} are the maximal distance difference and the maximal intensity difference respectively between p_m and p_n in the edge map, normalizing both two metrics at the same scale.

The optimal path ϕ^* can be found by minimizing the function defined as follows:

$$\begin{aligned} \phi^* \quad = \quad & \underset{\phi}{argmin} \quad \frac{1}{l_\phi}\int_{p_m, p_n \in \phi} c(\theta, p_m, p_n)d\theta \\ & \theta \in [0, 2\pi] \end{aligned} \tag{2.3}$$

where l_ϕ denotes the length for the path ϕ.

In the discrete setting, the function above can be rewritten as below:

$$\phi^* \quad = \quad \underset{\phi}{argmin} \quad \frac{1}{n_\phi}\sum_{\theta=0}^{2\pi} c(\theta, p_m, p_n) \tag{2.4}$$

where n_ϕ is the number of edge pixels along the path ϕ.

Dijkstra's algorithm is an algorithm widely applied to find the shortest path between the source node and the terminal node in a graph, such as road networks, telephone network and so on. It was first conceived in 1956 by Edsger W. Dijkstra [22]. In our case, Dijkstra's algorithm is naturally applied to solve the above minimization problem. We represent the edge pixels in

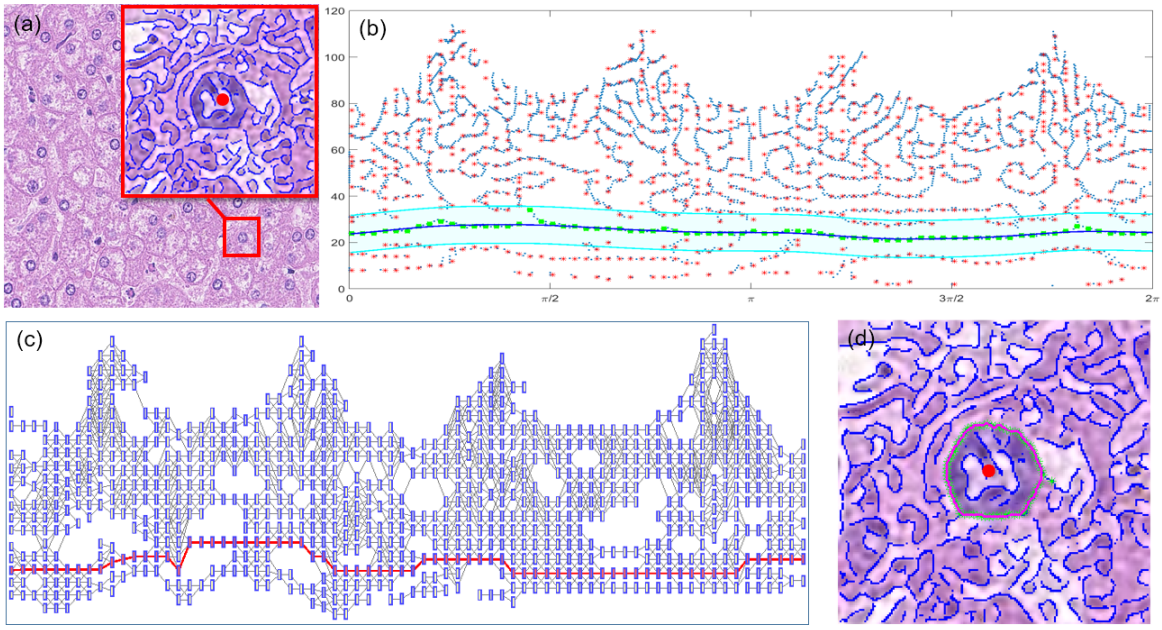

Figure 2.4: (a) Original image with a sub window showing the edge map detected by Canny edge detector for one particular nucleus ($\sigma_i = 3$) with the seed in the center (red dot). (b) Edge pixels are transformed into polar space with the nucleus seed being the coordinate origin. Red points are the locations with locally maximal number of pixels; Green points show the edge pixels along the optimal path searched by Dijkstra's algorithm. The blue solid line is the fitted curve. In the cyan area, edge pixels from the *i+1*th level are chosen as candidates. (c) Constructed undirected graph with nodes being the red points in (b) and edge weights being the cost defined by the combination of distance and intensity metrics. Nodes marked as red constitute the optimal path. (d) Final contour (red) and optimal path (green) are shown in the image patch.

polar space in the form of undirected graph $G = [V, E]$, where nodes in set V denote the edge pixels and set E denotes connections between any two adjacent angle nodes p_m, p_n with weights $c(\theta, p_m, p_n)$.

To make the algorithm robust against noisy structures, points in polar space are further represented/discretized by finding the locations with locally maximal number of edge pixels at each angle. A sliding window with width w and height h is constructed and move along the distance direction at each θ to capture and count the edge pixels locally. The number of edge pixels within the sliding window centered at $[r_j, \theta_i]$ is denoted as $N(r_j, \theta_i)$ and for angle θ_i locations with locally maximal number of pixels are denoted as $r_{l_{max}}(\theta_i)$. The detected locations are the discrete representations of original edge pixels in polar space (red dots in Figure 2.4(b)). Such approximations for edge pixels help reduce the number of possible paths connecting the source node and the terminal node and thus reduce the computational cost dramatically when searching for the optimal path using Dijkstra's algorithm.

The nuclei segmentation performance depends on the selection of source node and terminal node along the nucleus border. In the experiment, we propose to choose the source-terminal pairs as points with similar distances away from the seed at angle 0 and 2π in the graph. However, multiple source-terminal pairs may exist, and thus multiple optimal paths can be found by Dijkstra's algorithm. We note this on the edge map using a small σ where many noisy edge pixels with similar distances at the angle 0 and 2π can be source-terminal candidates. When irrelevant border pixels are selected as the source or terminal node, the optimal path would be searched by Dijkstra's algorithm at a high cost of passing through connections with large weights in the graph. The real nucleus contour is the route with the minimal cost connecting source-terminal pairs.

Some pixels along the optimal path are not necessarily on the real nucleus contour due to the incomplete border edges shown in the map, especially at locations with blurry boundries. Here, we apply the RANSAC [28] algorithm and the spline curve fitting method to estimate the nucleus contour locations. Given the edge pixels on the optimal path, a subset of pixels

are selected to generate a fitted curve model describing the rough shape of the optimal path. The points fitting well to the estimated model are called inliers and points with large estimation errors are called outliers. Afterwards, the model is refined using the inliers only. Such optimization process repeats for a fixed number of times and the model with the maximal number of inliers is considered as the reliable contour estimation. The curve fitting operation takes as input the inliers along the optimal path and outputs the final smooth contour C^k connecting the isolated and imcomplete edge segments when performed at the *k*th level of the edge pyramid.

Edge iteration

The smooth contour generated from the *k*th level is taken as the initial nucleus border and helps guide the edge selection for the *k+1*th level. Specifically, at the *k+1*th level, edge pixels within the distance range $[C^k - d, C^k + d]$ are chosen as edge candidates and edge pixels outside the range are discarded in the sense that a more refined contour at the *k+1*th level should be close to C^k with a distance tolerance d. When the blur parameter σ changes slightly, the edge locations change smoothly and would not shift much. Therefore, for the *k+1*th level, the edge pixel locations at angle θ should be within the range $[C^k(\theta) - d, C^k(\theta) + d]$. With the set of pixel candidates, the edge selection is performed as described above to generate a more accurate nucleus contour C^{k+1}.

As the contour is refined iteratively from the bottom level of the edge pyramid up to the top level, it gradually attaches to the real border of nucleus. Using a small blurring σ_0 at the top level, our algorithm can delineate the nucleus spatial content precisely.

2.3 Experimental results

2.3.1 Datasets

Tissue blocks and cytology slides were obtained from the archives of a local hospital (approved as an exempt protocol by the Institutional Review Board). Cases for analysis included liver

resection specimens and cytology slides prepared from fine needle aspiration biopsies of thyroid nodules.

Tissue procurement and processing

Liver tissues were procured at the time of a designated surgical procedure. All tissues were fixed in 10% neutral buffered formalin and processed on a conventional tissue processor using a series of graded alcohols and xylenes prior to paraffin embedding. Tissue sections were cut at 5 micron thickness from the paraffin-embedded block and placed on conventional 25 mm $\times$ 75 mm $\times$ 1.0 mm Superfrost Plus microscope slides using Fisherbrand Superslip cover slips (50 mm$\times$24 mm$\times$0.17 mm; Fisher Scientific, Thermo Fisher Scientific, Inc., Waltham, MA). All tissue sections for imaging were stained using conventional hematoxylin and eosin protocol used in the histology laboratory. For the thyroid cytology preparations, aspirate smears were fixed in 95% ethanol and then stained with the Papanicolaou (Pap) staining technique. Briefly, the Pap stain uses hematoxylin, OG-6, and eosin azure (combination of Eosin Y, Light Green SF, and Green FCF dyes) to stain cytological preparations. Nuclei stained with this technique have a blue-green color and excellent chromatin detail that can be visualized by light microscopy.

Digital image acquisition

Whole slide digital images of the liver slides were acquired using an Omnyx VL4 digital whole slide scanner (Omnyx, LLC) equipped with a 60$\times$ dry objective. Images obtained had a resolution of 0.1375 microns/pixel and were saved in the proprietary format then converted to lossless JPEG format. All thyroid cytology slide images were acquired using an Olympus BX51 microscope equipped with a 100$\times$ UIS2 UPlanFl oil immersion objective (numerical aperture 1.30; Olympus America, Central Valley, PA) and 2 megapixel SPOT Insight camera (Diagnostic Instruments, Sterling Heights, MI). Image specifications were 24 bit RGB channels and 0.074 microns/pixel, 118$\times$89 μm field of view.

2.3.2 Qualitative analysis

Before our algorithm is applied to pathology images, the nuclei channel should be extracted from RGB color space by color deconvolution [74] (*e.g.* extracting Hematoxylin channel from H& E stained images). After that, all image data is normalized to fit the intensity range [0, 1]. We tested the proposed method on two real datasets including thyroid dataset (35 representative images, Papanicolaou stained, 903 cell nuclei) and liver datasets (25 images, H&E stained, 2145 cell nuclei).

In our experiment, the sliding window width w and height h were set as 15 degrees and 2 pixels respectively and were fixed for both two datasets. The only parameter required to be changed for the two datasets is the maximal smooth level σ_{max} which was set to be 3 and 5 for the thyroid dataset and the liver dataset respectively. The minimal smooth level σ_{min} was set to be 1 in order to capture the precise nucleus border.

For comparison, we chose the following state-of-art algorithms for nuclei segmentation including the Ovuscule [80], level set [53] and template matching [12]. Template matching has the ability of both nuclei detection and segmentation while level set and the Ovuscule need predefined nuclei seeds for segmentation. In our experiment, level set and the Ovuscule adopted the seeds detected by MESPS to evaluate the segmentation performances.

For qualitative comparison, sample segmentation results by different approaches are shown in Figure 2.5, where the rows from the top to the bottom correspond to the results from level set, the Ovuscule, template matching and our method respectively and the columns from left to right correspond to sample images from liver dataset and thyroid dataset respectively.

2.3.3 Quantitative analysis

In addition, we evaluated the nuclei detection efficiency of template matching and MESPS using the miss rate (MR) defined as follows:

$$miss\,rate\,(MR) = \frac{SA \cup SM - SA \cap SM}{SM} \times 100\% \tag{2.5}$$

where SA are the seeds detected by the algorithms, SM are the seeds selected manually. SA $\cup$ SM and SA $\cap$ SM are the number of seeds in the union set and the intersection set of SA and SM respectively.

The segmentation accuracy was measured by the area error rate (AER)[76] focusing on the number of incorrectly segmented pixels and the spatially-aware evaluation metric normalized sum of distances (NSD)[15] with the ground truth. Quantitative analysis of nuclei detection and segmentation efficiency of different approaches is shown in Table 2.1.

From the quantitative evaluations of different approaches, we note that the proposed method showed similar or superior performance compared with existing segmentation methods validated on two datasets. For the thyroid dataset, level set segmented nuclei with the highest accuracy with AER and NSD being 8.31% and 0.29 respectively. Our method generated similar results as that of level set, showing that MESPS achieves nuclei segmentation performance comparable to the state-of-art method. Moreover, for the liver dataset in the complex setting (nonuniform illumination, noisy background and nuclei heterogeneity), MESPS was still able to find the nuclei borders accurately and performed the best compared with the listed approaches. Considering the comprehensive performance over the two validation datasets, MESPS archived the best segmentation accuracy with 10.34% AER and 0.33 NSD on average.

Table 2.1: Quantitative evaluation of different approaches on nuclei detection and segmentation efficiency

Algorithms		Level set	The Ovuscule	Template matching	MESPS
AER/NSD ($\times$10)	Thyroid dataset	8.31%/0.29	12.63%/0.44	15.29%/0.46 MR: 29.24%	8.45%/0.31 MR: 27.97%
	Liver dataset	21.39%/0.64	17.46%/0.42	19.56%/0.44 MR: 21.19%	12.22%/0.35 MR: 26.21%

For the overall nuclei detection efficiency, both template matching and MESPS could detect

most manually labeled nuclei with the similar miss rates. However, we should be aware of the supervised fashion of template matching method that needs users to select a set of nuclei for training and then finds the templates within the constructed statistical model that best match the testing nuclei .

2.4 Conclusion

This chapter described an unsupervised method to detect and segment nuclei automatically from the 2D pathology images based on normalized cross correlation (NCC) and multi-scale edge selection in polar space. The experimental validations showed that the method could segment the nuclei accurately when applied to real pathology images with different stainings (*e.g.* H&E, Pap staining) and image qualities (*e.g.* blurring, noise, texture heterogeneity).

There are several advantages of the method. First, it has the ability to locate nuclei borders precisely with certain robustness. The multi-scale strategy ensures that the ill effects caused by noise, nonuniform intensity etc., are greatly reduced by smoothing. The small smooth level at the top of the edge pyramid make contour gradually cling to the real nucleus border at the pixel level. With the small step size of σ, the contour changes smoothly as it iterates from the bottom level up to the top level of the edge pyramid. Second, it is designed in an unsupervised way that doesn't need users to train a segmentation model. Once the parameters are set, the algorithm could detect and segment nuclei with little user effort. The performance of MESPS depends on the image gradient, thus it is not sensitive to staining techniques or imaging modalities, making it useful and applicable to various datasets in clinic settings. Finally, the proposed algorithm is light weight, consisting of several basic but effective algorithms including normalized cross correlation, edge detection and Dijkstra's algorithm. The proposed framework is mathematically simpler than the state-of-art approaches.

In addition, we proposed some ways to reduce the computational cost by reducing the number of nodes and edges in the graph. First, edge pixels in polar space are represented by the

points with locally maximal number of pixels within a sliding window at each angle. This operation reduces the number of nodes greatly in the graph with the additional benefit of denoising. Moreover, edges between two adjacent nodes in the graph would be deleted if the weight is over a certain threshold, in the sense that neighboring border pixels should be near each other and the distance away from the nucleus center changes little. Each node only connects few nodes at adjacent angles, which prominently cut down the number of possible paths between the source and terminal nodes.

We should note that there are some parameters introduced in our proposed method, including filter size, σ_{min}, σ_{max}, smooth step size, cost weights (a, b) and the threshold for edge deletion. Parameters to be changed for various datasets are filter size and σ_{max}. As described in the previous section, the filter size depends on the image resolution as well as the nuclei type in study. σ_{max} can be determined based on the image gradient in datasets, which is set to keep the correct nuclei edges in the edge map, and at the same time filter out the noisy edges. In practice, the optimal σ_{max} can be set experimentally by randomly selecting a few sample images in the dataset, and observing the edge maps so that the edge detector could describe the outlines of most nuclei. Our algorithm is not sensitive to other parameters and they were fixed when validated on two datasets. In our experiment, the values of σ_{min}, smooth step size, a,b and the threshold were set to be 1, 0.5, 0.4, 0.6 and 20 respectively.

Besides the advantages mentioned above, the method has some limitations that are noteworthy of discussion. First, the method is designed for segmenting convex shaped nuclei. In the polar space, parts of the contour for non-convex shaped nucleus are mapped to multiple locations at the same angle, which violates our assumption that there is only one optimal border location per angle. Even though most nuclei have the shape of sphere or ellipsoid, highly concave nuclei can be observed under microscopy due to the sectioning of nuclei at odd angles or tissue distortion in slides preparation procedure, or both. Second, the method can't handle overlapping nuclei even if the nuclei seeds are detected correctly. Due to the blurring within the nuclei overlapping area, the edge detector usually does not generate edges delineating the two nuclei. The method

would treat the two nuclei as one and produce the border of non-overlapping area. However, we should note that 1) the ultimate goal of nuclei segmentation is for exploring the correlation between nuclei morphology and cellular/ disease progress. 2) overlapping nuclei provides limited information for analysis due to the difficulty of recovering inherent information within the overlapping area. Therefore, with plenty of isolated nuclei available in the dataset, nuclei overlapping problem is negligible in subsequent nuclei analysis.

The proposed method locates nuclei by measuring the matching degrees between local image patches and the predefined filters. Afterwards, the method transforms the object segmentation problem into the shortest path problem in a graph. The cost function is constructed considering both shape and intensity characteristics of nuclei borders. The accurate delineation of nuclei is based on the detected border pixels which can be correctly selected by the well-known Dijkstra's algorithm. The multi-scale strategy enables the contour generated at each level evolves smoothly to the actual nucleus border. In the future, the method could be further automated by enabling the algorithm to select the optimal maximal smooth parameter based on image gradient statistics.

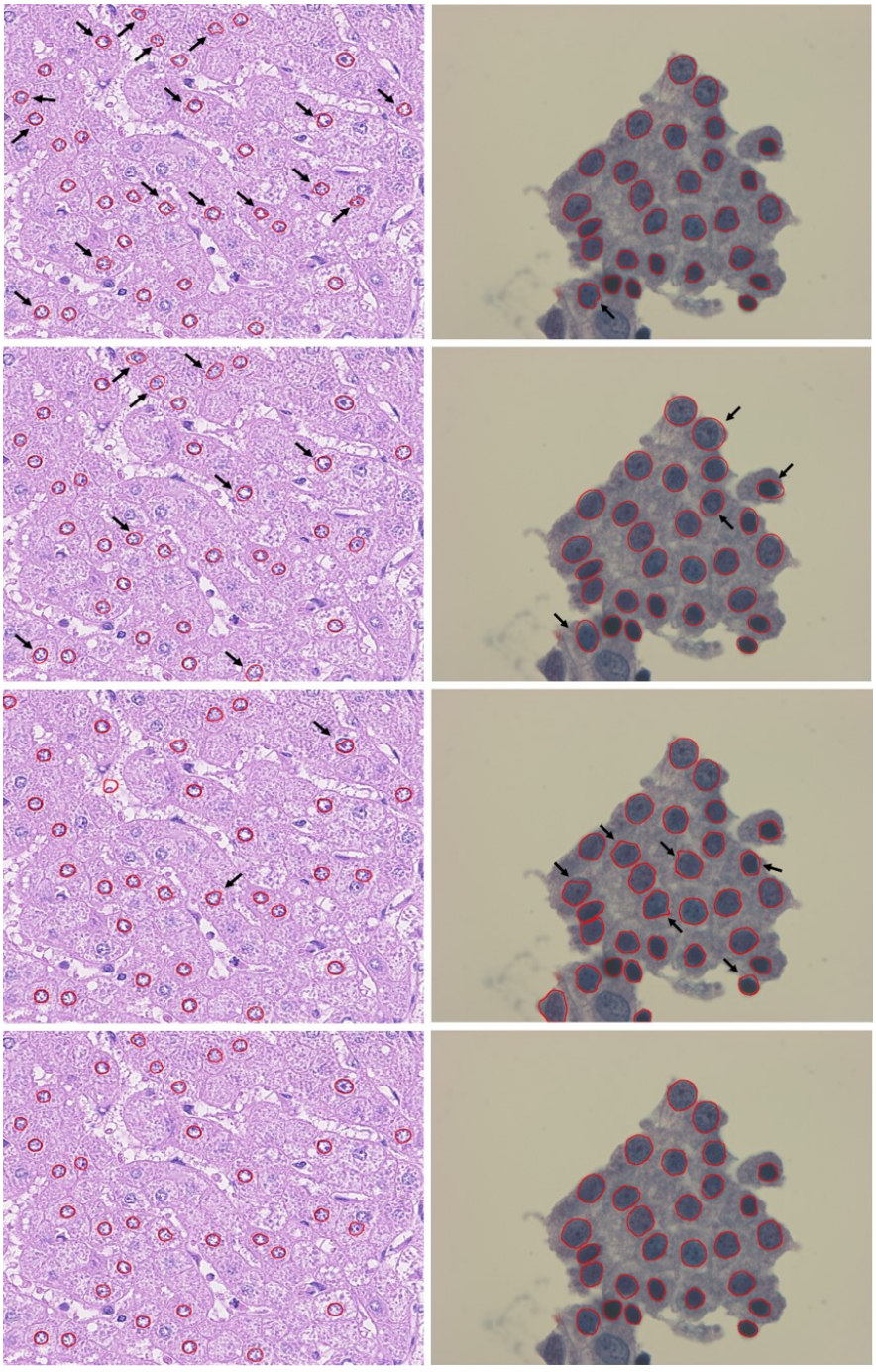

Figure 2.5: Segmentation results from two validation datasets. First column: liver dataset; second column: thyroid dataset. From the top row to the last row are the results by level set, the Ovuscule, template matching and MESPS respectively. Note that segmentation flaws are pointed out by black arrows.

Chapter 3

SetSVM: An Approach to Set Classification in Cancer Detection

3.1 Introduction

Due to the importance of nuclear structure in cancer diagnosis, several predictive models have been described for diagnosing a wide variety of cancers based on nuclear morphology. In many computer-aided diagnosis (CAD) systems, cancer detection tasks can be generally formulated as the set classification problem, which is different from single instances classification.

In clinical diagnosis, pathologists rely on microscopic examination of a set of nuclei within the tissue sample for analysis. Thus, in most situations, a diagnostic label is only available for the tissue sample rather than individual nuclei. A predictive model is required to learn from sets of nuclei without nuclei-level annotations and predict the diagnostic label for a new set of nuclei, referred as set classification problem. Beyond cancer diagnosis, set classification problem is also ubiquitous in prognosis prediction, where the model needs to predict the patients survival outcome by taking account of a set of quantified nuclei [96]. Different from conventional image classification where training and testing samples are labeled single-shot images, in the set classification scenario, training and testing samples are sets, each of which consists of various

numbers of unlabeled nucleus images. The set classification problem is challenging and can't directly solved by supervised machine learning approaches. Limitations of existing approaches to set classification are described as follows.

Single image classification (1.2.2) essentially seeks to build an instance-level classifier utilizing set-level labels, which infers the latent instance labels for set prediction. However, it is reasonable to expect that not all nuclei in the tissue sample show characteristic morphological changes associated with the disease. For example, besides cells that exhibit abnormal phenotypes, cancerous tissues contain cells displaying normal phenotypes. Therefore, the choice of assumption for single instance classification often requires prior domain knowledge and has significant impact on overall performance of the predictive model.

For existing approaches using the concept of set classification, we note two separate processes in these methods when dealing with set classification problems. First, a mapping function for set representation or set-wise distance metric is constructed in an unsupervised way. For example, in STATS, the types of statistics are manually defined; in BoW, the dictionary for histogram representation is built by k-means algorithm with the cost function aiming at minimizing reconstruction error. Second, a set-level classifier takes as input set representations for supervised training based on certain criterion, e.g. minimizing classification error or maximizing separation margin. Although existing set classification approaches have achieved different degrees of success in cancer prediction, a shared problem is that the model performance may be limited due to the inconsistent objectives in two separate processes.

In this chapter, we propose a novel set classification method, SetSVM, which unifies set representation learning with classifier training. The method solves set classification problem by jointly optimizing the mapping function and the SVM decision boundary in a maximum soft margin problem. We show that a better performance is possible by introducing discriminant information from the classifier to the mapping function. Beyond cancer detection, we use invertible features to show that SetSVM is able to visualize set-level morphological attributes in a discriminative subspace, which helps interpret patients nuclear patterns from different classes.

We test the effectiveness of SetSVM using different types of nuclear quantification approaches, provide comparisons with five state-of-art methods and provide experimental validations with 260 patients in total in four diagnostic challenges.

3.2 Method

Suppose we have N patients included for study, the nuclei dataset can be denoted as tuples $X = \{(X_i, y_i), i = 1, ..., N\}$, where $X_i = \{x_{ia}\}_{a=1}^{n_i}$ is a set of nuclei extracted from the *i*th patient and $y_i \in \{-1, +1\}$ (normal vs. cancer) is the corresponding patient label. One single nucleus is quantified as $x_{ia} \in \mathbb{R}^{d\times 1}$ describing its morphological characteristics. The goal of set classification is to find class label y_{test} to which the unseen set X_{test} belongs.

A collection of prototypes are firstly initialized (section 3.2.3), SetSVM then constructs a mapping function (section 3.2.1) to extract set representation to describe the global nuclear attributes for any nuclei set. The idea is to compute matched nucleus in a set with respect to each prototype. In model training (section 3.2.2), both prototypes and decision boundary are jointly optimized to maximize the separation margin, leading to discriminative set representations for specific cancer detection tasks. In this chapter, we show that the optimization of prototypes is a mathematical variant of the learning vector quantization (LVQ) technique (section 3.2.4).

3.2.1 Nuclei Set Representation via Prototypes

The mapping function for set representation stems from the well-known matching kernel in pattern recognition [84]. The basic idea of matching kernel is to measure the maximal similarity between two sets X_i and X_j by finding the matched instance in one set with respect to a particular instance in the other set.

The matching kernel can be defined as follows:

$$K(X_i, X_j) = \sum_{a=1}^{n_i} \underset{b=1,\ldots,n_j}{max}\ k(x_{ia}, x_{jb}) + \sum_{b=1}^{n_j} \underset{a=1,\ldots,n_i}{max}\ k(x_{ia}, x_{jb}) \tag{3.1}$$

where k is a local Mercer kernel for instances from two sets. K is computed by finding the matched instance in X_j for any $x_{ia} \in X_i$ and the matched instance in X_i for any $x_{jb} \in X_j$. However, due to the max operation, K is not necessarily positive definite even if k is a Mercer kernel and therefore it is risky to use it as SVM kernel for set classification.

The kernel can be positive definitive by comparing matched pairs of instances in X_i and X_j via a collection of prototypes, defined as follows:

$$K(X_i, X_j) = \sum_{p_k \in P} k(\Phi(X_i), \Phi(X_j)) \tag{3.2}$$

where $P = \{p_k\}_{k=1}^{m}$ is a set of prototypes containing m pre-defined prototypes; $\Phi_{p_k}(X_i)$ is to find matched instance in X_i by prototype p_k; k is generally defined as a positive definite kernel (e.g. linear kernel, radial basis function, polynomial kernel), which makes K positive definite.

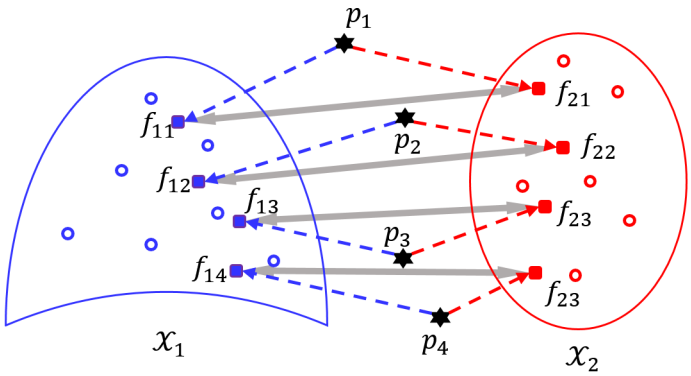

Figure 3.1: Illustration of the mapping function via prototypes. Matched instances (marked in red and blue squares) in X_1 and X_2 are computed regarding each prototype and are used for set representation. Here we have four prototypes: $P = \{p_1, p_2, p_3, p_4\}$.

Here, we define $\Phi_{p_k}(X_i)$ as the weighted mean of instances in X_i related to p_k:

$$\Phi_{p_k}(X_i) = \frac{\sum_{x_{ia} \in x_i} h_{p_k}(x_{ia}) x_{ia}}{\sum_{x_{ia} \in x_i} h_{p_k}(x_{ia})} \tag{3.3}$$

where $h_{p_k} = e^{-\gamma \|x_{ia} - p_k\|_2^2}$ using the RBF kernel and $\Phi_{p_k}(X_i)$ outputs the mapping in the same feature space as instances. $h_{p_k}(x_{ia}) \in [0, 1]$ measures the matching degree between instance x_{ia} and prototype p_k. γ is a smoothing parameter and $\Phi_{p_k}(X_i)$ in equation 3.3 equals finding the nearest neighbor in X_i for p_k when γ is sufficiently large. Fig. 3.1 presents the construction of the mapping $\Phi_{p_k}(X_i)$.

We choose k as the linear kernel, thus equation (3.2) becomes:

$$\begin{aligned} K(X_i, X_j) &= \sum_{p_k \in P} \Phi_{p_k}^T(X_i) \Phi_{p_k}(X_j) = f_i^T f_j \\ f_i &= f_{i1} \circ \cdots \circ f_{ik} \cdots f_{im} \end{aligned} \tag{3.4}$$

where $f_{ik} = \Phi_{p_k}(X_i), k = 1, ..., m$ and $f_i \in \mathbb{R}^{md \times 1}$ is the set representation concatenated by the mapping $\Phi_{p_k}(X_i)$ with respect to prototype p_k, summarizing the instance attributes in X_i. The kernel defined in equation (3.4) not only measures the similarity between two sets X_i and X_j, it also explicitly provides set representation for any nuclei set, facilitating visualization of discriminative information in the nuclei set as will be seen in section 3.3.3.

3.2.2 Unifying Set Representation Learning with Classifier Training

As mentioned earlier, the uniqueness of SetSVM is to combine set representation learning with classifier training in one unified cost function to increase discriminativeness. Here we maximize the soft separation margin over both the decision boundary and the prototypes P. In a two-class classification problem, SVM classifier seeks to find a hyperplane: $wx + b = 0$, which maximizes the separation margin. SVM can be formulated as the following unconstrained minimization problem with hinge loss:

$$L(\theta) = C\|w\|_2^2 + \frac{1}{N}\sum_{i=1}^{N} l(y_i(w\phi(x_i) + b)) \tag{3.5}$$

where $\theta = \{w, b\}$ is the SVM decision boundary, $l(u) = max(0, 1 - u)$ is the hinge loss term, $\phi(\cdot)$ is a kernel function mapping x_i to a high dimensional space, C is a parameter for the regularization term. One point x can be classified by $sgn(\sum_{i=1}^{N} \alpha_i y_i \phi(x_i, x) + b)$ with α_i being the Lagrange multipliers.

For set classification, the kernel function $\phi(x_i, x)$ is replaced by $K(X_i, X)$ in equation (3.4). The class label for a test set X is determined by $sgn(\sum_{i=1}^{N} \alpha_i y_i K(X_i, X) + b)$. As mentioned in the previous section, we choose k in equation (3.4) as linear kernel and thus get the linear SVM taking as input set representations f_i and f_j for X_i and X_j respectively. Denote 3.5 as $L(\theta, P)$, the optimization process is performed in two phases: 1) fix prototypes P, optimize $L(\theta, P)$ over θ. Since the problem is convex, the optimal decision boundary is feasible by quadratic programming (QP) algorithms. 2) fix the decision boundary θ, optimize prototypes P. Due to the non-convexity of cost function, the optimal P can be learned by gradient descent based approach. Such optimization processes proceed alternatively.

Although the update of P is straightforward to derive, we provide the details here for completeness and for showing the relation between our method and LVQ (learning vector quantization) in the next section.

$$\begin{aligned} \nabla_{p_k} L &= \frac{1}{N}\sum_{i=1}^{N} \nabla_{p_k} l_i \\ l_i &= max(0, 1 - y_i(w^T f_i + b)) \end{aligned} \tag{3.6}$$

Since the hinge loss in L is not differentiable, sub-gradient is computed. When $y_i(w f_i + b) < 1$, the derivatives of $L_i(X_i;\ \theta, P)$ with respect to each $p_k \in P$ are as follows:

$$\nabla_{p_k} l_i = \nabla_{p_k} f_{ik} l_i \tag{3.7}$$

$$\nabla_{f_{ik}} = -y_i w_k \in \mathbb{R}^{d\times 1}, w = w_1 \circ \cdots \circ w_k \circ \cdots w_m \tag{3.8}$$

$$\nabla_{p_k} f_{ik} = [\nabla_{p_k} f_{ik_1} \circ \cdots \nabla_{p_k} f_{ik_l} \circ \cdots \nabla_{p_k} f_{ik_d}] \in \mathbb{R}^{d\times d} \tag{3.9}$$

$$\nabla_{p_k} f_{ik_l} = 2\gamma \frac{\sum_{x_{ia}\in X_i}(x_{ia} - p_k) h_{p_k}(x_{ia})(x_{ia_l} - f_{k_l})}{\sum_{x_{ia}\in X_i} h_{p_k}(x_{ia})}$$

When $y_i(wf_i + b) \geq 1$, the sub-gradient $\nabla_(p_k) l_i = 0$ for all prototypes. Thus the derivative $\nabla_(p_k) L_i$ can be organized as:

$$\begin{aligned} \nabla_{p_k} L &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{\alpha_i > 0\} \nabla_{p_k} f_{ik} \nabla_{f_{ik}} l_i \\ &= -\frac{2\gamma}{N}\sum_{i=1}^{N} \mathbb{I}(\alpha_i > 0) y_i \frac{\sum_{x_{ia}\in X_i} h_{p_k}(x_{ia})(x_{ia} - p_k) u_{ia}^k}{\sum_{x_{ia} in X_i} h_{p_k}(x_{ia})} \end{aligned} \tag{3.10}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function; $u_{ia}^k = w_k^T(x_{ia} - f_{ik}) \in \mathbb{R}$.

We should note that the Lagrange multipliers $\alpha_i = 0$ if $y_i(wf_i + b) \geq 1$ and $\alpha_i > 0$ if $y_i(wf_i + b) \leq 1$, meaning that only support vectors contribute to the updates of prototypes. With learning rate $\lambda > 0$, the prototype p_k is updated by:

$$p_k^{t+1} = p_k^t - \lambda \nabla_{p_k} L; k = 1, ..., m \tag{3.11}$$

3.2.3 Initialization of Prototypes

Prototypes P are initialized with the clustering method across all nuclei in the training data. To make the initialization procedure robust, we apply *k*-means++, where the first cluster center is chosen randomly and each subsequent cluster center is chosen from instances with probability proportional to its squared distance from the point's closest existing cluster center.

The number of clusters m is predefined. We note that the initialization of prototypes with *k*-means clustering is unsupervised and is not necessarily optimal for classification purposes.

Algorithm 1 Proposed Set Classification Approach SetSVM

Input: $X = \{X_i\}_{i=1}^{N},\ y = \{y_i\}_{i=1}^{N},\ \gamma,\ m$

Output: $P = \{p_k\}_{k=1}^{m},\ \theta = \{w, b\},\ f_i$

1. Initialize prototypes P with *k*-means clustering
2. Repeat alternative prototype and classifier learning
3. Fix $P, \theta = \underset{\theta}{argmin}\ L(P, \theta)$
4. Fix θ, for $k = 1$ to m do
5. $\quad p_k^{(t+1)} = p_k^t - \lambda \Delta_{p_k} L(P, \theta)$
6. end for
7. $F(X_i|P) \rightarrow f_i$

3.2.4 Relation to LVQ

Learning vector quantization (LVQ) [48] is a supervised prototype learning method using class label information for pattern recognition tasks. The set of prototypes are defined in the feature space of the observed data. The essence of LVQ is that one prototype p_k is updated toward the direction of data point x if they have the same class labels; otherwise the prototype is repelled, which can be defined as follows:

$$p_k \leftarrow \begin{cases} p_k + \lambda(x - p_k), c(p_k) = c(x) \\ p_k - \lambda(x - p_k), c(p_k) \neq c(x) \end{cases} \tag{3.12}$$

where $c(x)$ and $c(p_k)$ are the class labels for x and p_k respectively.

In our model, the update rule for p_k in equation (3.11) can be written as:

$$p_k \leftarrow p_k + \frac{2\lambda\gamma}{N}\sum_{i=1}^{N}\mathbb{I}(\alpha_i > 0)y_i\frac{\sum_{x_{ia}\in X_i} h_{p_k}(x_{ia})(x_{ia}-p_k)u_{ia}^k}{\sum_{x_{ia}\in X_i} h_{p_k}(x_{ia})} \tag{3.13}$$

For simplicity, let's consider X_i with $\alpha_i \neq 0$:

$$p_k \leftarrow p_k + 2\lambda\gamma\frac{\sum_{x_{ia}\in X_i} h_{pk}(x_{ia})(x_{ia}-p_k)y_i w_k^T\Delta}{\sum_{x_{ia}\in X_i} h_{p_k}(x_{ia})} \tag{3.14}$$

where $\Delta = x_{ia} - f_{ik}$ is the change for f_{ik} along the direction toward x_{ia}. During model learning, the contribution of set X_i on the update of prototype p_k is a weighted contribution of all instances belonging to X_i. The update direction of p_k based on instance x_{ia} is determined by $sgn(y_i w_k^T\Delta)$.

If $y_i w_k^T\Delta > 0$, the change Δ for f_{ik} is predictive about the set class label y_i and prototype p_k will be pulled toward instance x_{ia} to narrow the gap $\|x_{ia} - p_k\|_2$, leading to a higher weight $h_k(x_{ia})$ for x_{ia} in generating f_{ik} in equation (3.3). In contrast, if $y_i w_k^T\Delta < 0$, the change Δ based on x_{ia} is not predictive and thus the prototype p_k will be repelled, leading to a smaller weight $h_(p_k)(x_{ia})$ for x_{ia} in f_{ik}. As a result, the gradient-based updates of prototypes P in SetSVM can be viewed as a relaxed version of LVQ utilizing set-level label information.

3.3 Experiments

3.3.1 Datasets

Diagnostic challenges in thyroid cancer, liver cancer and melanoma were included in this study for quantitative evaluation of the proposed SetSVM in cancer detection tasks. Under an Institutional Review Board approval, tissue blocks for the thyroid and liver datasets were obtained from the archives of the University of Pittsburgh Medical Center (UPMC). Tissue blocks for the melanoma dataset were retrieved from the archives of the Mount Sinai Hospital. All cases were reviewed by more than one pathologist, and only cases with a clear diagnosis (gold standard) were selected for this study.

In the thyroid dataset, cases for analysis included resection specimens with diagnosis of three

different types of thyroid lesions, namely follicular adenoma of the thyroid (FA), follicular variant of papillary thyroid carcinoma (FVPC), and nodular goiter (NG). All tissues sections were stained with Feulgen technique. 78 patients were involved in the thyroid dataset with 28 patients for FA, 22 patients for FVPC and 28 patients for NG. The diagnostic challenges in the thyroid dataset included differentiating FA from NG as well as differentiating FVPC from NG.

In the liver dataset, tissue sections for imaging were stained using conventional hematoxylin and eosin (H&E) protocol used in the histology laboratory. Cases for analysis included 26 specimens with diagnosis of focal nodular hyperplasia (FNH) and 17 specimens with diagnosis of malignant hepatocellular carcinoma (HCC). The diagnostic challenge included differentiating FNH from HCC.

In the melanoma dataset, tissue sections were stained using H&E. A total of 139 cases were included in our study, including 67 cases diagnosed with malignant melanoma (MM) and 72 cases diagnosed with dysplastic nevi (DN). The diagnostic challenge included differentiating MM from DN.

3.3.2 Nuclei Segmentation and Preprocessing

Due to the differences in nuclei heterogeneity, cell nuclei in the three datasets were segmented with separate approaches. In the thyroid dataset, nuclei were segmented with a supervised method [12] while nuclei in the liver dataset and melanoma dataset were segmented using the unsupervised method described in 2.

Segmented nuclei images were preprocessed as follows. Each nucleus image I with intensities in the range (min,max) is first normalized by $(I - min)/(max - min)$ to minimize the intensity variations in slide preparation, staining procedure and image acquisition. Next, nuclei position variations such as rotation, translation and coordinate inversions are eliminated by position normalization. Nuclei are relocated to the image centers to remove translation and the major axes for nuclei in each dataset are aligned in the same direction to eliminate arbitrary rotation. A few segmented nuclei after preprocessing are shown in Fig. 3.2.

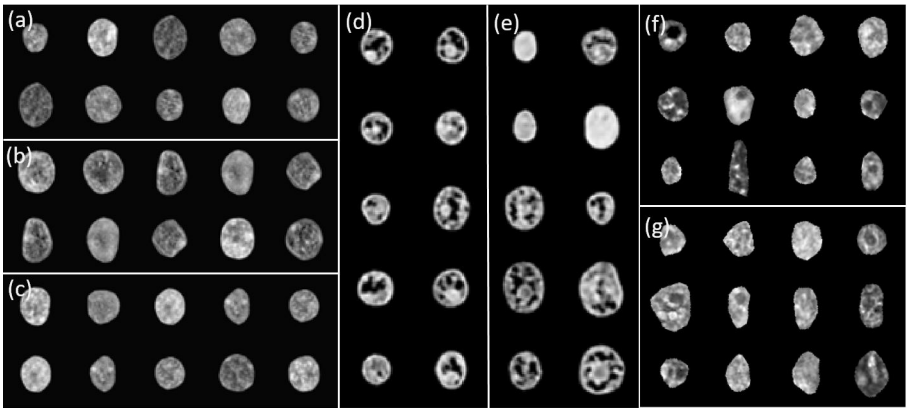

Figure 3.2: Segmented nuclei randomly selected from patients diagnosed with FA (a), FVPC (b), NG (c), FNH (d), HCC (e), DN (f) and MM (g). Nuclei intensity and position are normalized as described in section 3.3.2.

3.3.3 Nuclear Morphometry Quantifications

In the experiments, we used hand-crafted features, autoencoder features and transport-based morphometry to describe nuclear morphology and validate the effectiveness of SetSVM. All features are extracted in unsupervised way.

Hand-crafted features

Nuclear structural characteristics were quantified with 256-dimensional hand-crafted features, including 6 morphological features (*e.g.* area, perimeter, circularity), 220 texture features (e.g. Haralick features, Gabor features) and 30 wavelet features. Features were normalized to have the same variances of one.

Autoencoder features

Nuclear morphology was quantified by hidden features in a two-layer sparse stacked autoencoder (SSAE) [93]. The SSAE is able to transform nuclear quantification back to image space by input reconstruction, making it possible to visually investigate the feature space.

The SSAE transforms input image I_i to feature representation H_i in the hidden layer by the encoding phase: $T(I_i) \rightarrow H_i$, which is then used to approximate the input in decoding phase:

$T^{-1}(H_i) \rightarrow \hat{I}_i$. Since SSAE is stacked by basic SAE in a repeated fashion, we only provide a brief review for the single layer SAE. The basic SAE transforms the input image I_i into a new feature representation H_i in the hidden layer by the encoding phase: $T(I_i) \rightarrow H_i$, which is then used to approximate the input in decoding phase: $T^{-1}(H_i) \rightarrow \hat{I}_i$. The SAE model is optimized by minimizing the following problem:

$$J_{sparse} = \frac{1}{N}\sum_{i=1}^{N} J(I_i|W,B) + \beta_1 \sum_{k=1}^{s} KL(\rho\|\hat{\rho}_k) + \beta_2\|W\|_2^2 \tag{3.15}$$

where W, B are SAE parameters; $J(I_i|W,B) = \frac{1}{2}\|\hat{I}_i - I_i\|_2^2$ is the term for minimizing reconstruction error; $KL(\rho\|\hat{\rho}_k)$ is the Kull-Leibler (KL) divergence between the average activation of the k^{th} hidden unit (s hidden units in total) and the desired activation ρ; $\|W\|_2^2\|$ is the regularization for weights W; β_1 and β_2 controls the importance for corresponding terms. Weights W and bias B can be optimized by gradient descent approach.

In the experiment, we set the number of the first layer hidden units as 400 and the number of the second layer hidden units as 49. The visualization of learned weights in the two-layer SSAE is shown in Fig. 3.3.

Transport-based morphometry

Transport-based morphometry is an approach to quantify nuclear structure, refer to [87] for details. Briefly, nuclear image I_i is approximated by M particles using weighted kmeans method to summarize pixel intensity distribution over image coordinates. I_i can thus be denoted as $I_i = \sum_{p=1}^{M} m_p \delta_{x_p}$, where $\delta_{x_p}$is the delta function placed at position x_p and m_p is the corresponding intensity mass. Reference image I_0 is usually the average image across the dataset and can be similarly approximated by $I_0 = \sum_{q=1}^{M} m_q \delta_{y_q}$. Image approximations for I_0 and I_i are shown in Figure 3.4 (a) and (b) respectively. We seek to find the optimal transport plan T by minimizing the transport distance $d(I_0, I_i)$ between I_0 and I_i:

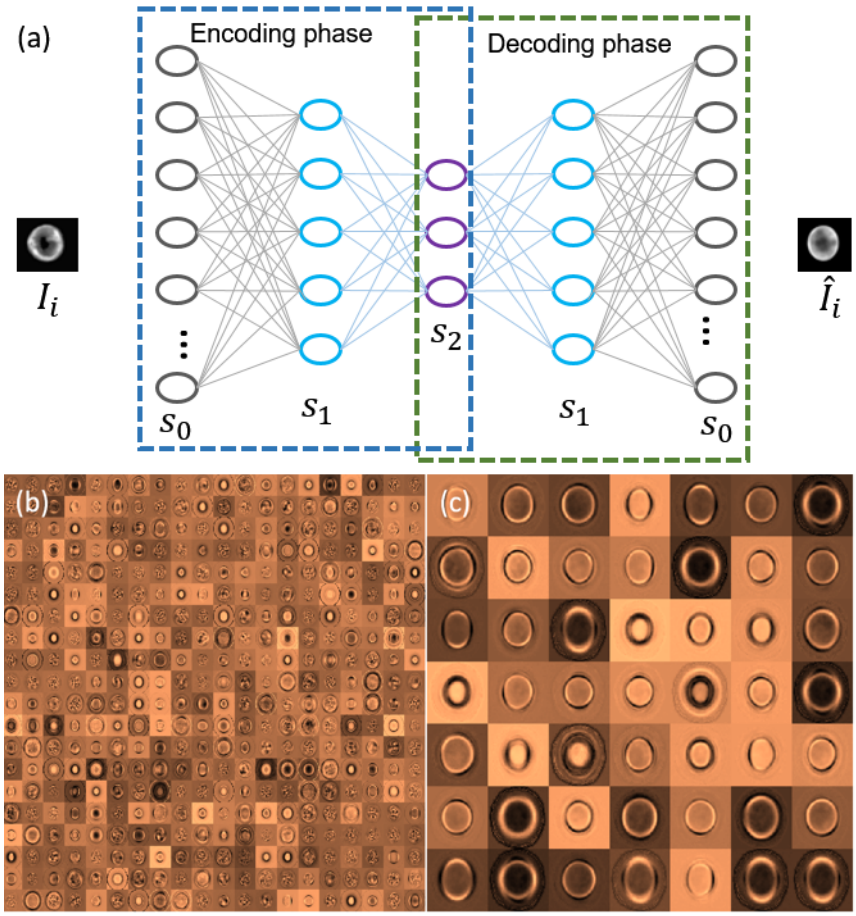

Figure 3.3: (a) Architecture of the two-layer stacked sparse autoencoder (SSAE). (b) Learned weights (20×20) in the first layer. (c) Learned weights (7×7) in the second layer. The grayscale images (b) and (c) were color coded for viewing purpose.

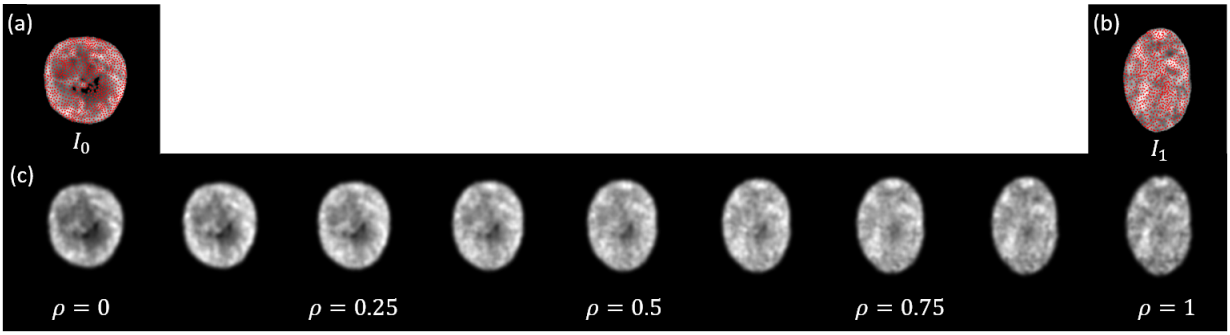

Figure 3.4: The red dots in I_0 (a) and I_i (b) are locations of particle masses to approximate each image. (c) shows visual representations along the geodesic path from I_0 to I_i.

$$d(I_0, d_i) = \min_{T} \sum_{p=1}^{M} \sum_{q=1}^{M} \|x_p - y_q\|_2^2 T_{pq},\ T \in \mathbb{R}^{M \times N} \tag{3.16}$$

The morphological representation U_i for I_i is obtained by:

$$U_i = [\sqrt{q_1} e_1, ..., \sqrt{q_M} e_M],\ e_q = \sum_{p=1}^{M} x_p T_{pq} / m_q \tag{3.17}$$

The geodesic interpolation between I_0 and I_i can be approximated by:

$$I_\rho = \sum_{j=1}^{M} q_j \delta_{pe_j + (1-\rho) y_j},\ \rho \in [0, 1] \tag{3.18}$$

Figure 3.4 (c) shows the visualization of the morphing processfrom I_0 to I_i when changing ρ from 0 to 1. The linear embedding preserves the information to approximate each image I_i, facilitating visualization and quantitative analysis of nucleus morphology at the same time.

Cross Validation

We utilized the leave-one-out strategy to test the cancer detection performance of SetSVM. The data from one patient is used for testing and the remaining data is used for model training. The training data was further split into training and validation sets to search for the best parameter γ in equation (3.3) and the number of prototypes m in SetSVM.

It is worth mentioning that testing patients were not involved in SSAE model training. We used standalone data to train SSAE just once for each diagnostic challenge in order to avoid multiple model trainings with leave-one-out strategy. That is, for challenge FA vs. NG, the standalone data is FVPC; for challenge FVPC vs. NG, the standalone data is FA; for challenge FNH vs. HCC, the standalone data is the thyroid data and for MM vs. DN, the standalone data is the liver data. After SSAE training, the model can be viewed as a feature extractor applied to both training and testing sets in the same manner to describe nuclear structure.

Visualizing Nuclear Structure Differences Between Sets

As mentioned earlier, set representation by SetSVM enables visualization of relevant differences between sets in terms of nuclear morphometry. Such visualization of discriminative information is plotted at patient level rather than nuclei level, reflecting characteristic nuclear morphology for the entire set.

We begin by finding the most important prototype in P for a two-class diagnostic challenge. Each prototype $p_k \in P$ maps set X_i into f_{ik} and the set representation for X_i is thus $f_i = f_{i1} \circ \cdots f_{ik} \cdots \circ f_{im}$. The optimal SVM decision boundary $w = w_1 \circ \cdots w_k \cdots \circ w_m \in \mathbb{R}^{md \times 1}$ indicates the importances of feature variables. We find the most important prototype $p_o \in P$ regarding differentiating nuclei sets by feature ranking:

$$o = \underset{k=1,\ldots,m}{argmax} \frac{\|w_k\|_2^2}{\|w\|_2^2} \tag{3.19}$$

Combined with the penalized version of Fisher Linear Discriminant Analysis (pLDA) [86], the set information f_{io} extracted by p_o can be used for visualizing discriminant variations between nuclei sets.

We have N data representations $f_{io} \in \mathbb{R}^{d \times 1} \mid i = 1, ..., N$ with each index i belonging to class c. The $pLDA$ seeks to find a discriminative direction V_{pLDA} by optimizing the following problem:

$$V_{pLDA} = \underset{V}{argmax} \frac{V^T S_T V}{V^T (S_W + \epsilon I) V} \tag{3.20}$$

where $S_T = \sum_i (f_{io} - \bar{f}_o)(f_{io} - \bar{f}_o)^T$ is the total scatter matrix; $S_W = \sum_c \sum_{i \in C} (f_{io} - \bar{f}_o)(f_{io} - \bar{f}_o)^T$ is the within class scatter matrix; $\bar{f}_o$ is the center of the entire dataset and $\bar{f}_c$ is the center for class c; ϵ is a constant.

We can project data points onto the discriminant direction V_{pLDA} by $V_{pLDA}^T f_{io}$. The discriminant variations along V_{pLDA} can be computed about the mean:

$$f_\mu \ = \ \bar{f}_o + \mu V_{pLDA} \tag{3.21}$$

where μ is the coefficient computed in units of the standard deviation σ for data projections along V_{pLDA}.

Note that if data samples are measured with invertible feature transformation $T(\cdot)$, f_μ can be transformed back to the image space by $T^{-1}(f_\mu)$. In our case, the decoding phase in SSAE can be used to visualize such variations by image reconstruction.

3.4 Results

3.4.1 Classification Accuracy Comparisons

To evaluate cancer detection performance of SetSVM, we also tested five existing approaches in the diagnostic challenges. All the methods take as input the same nuclear quantifications and utilize the SVM classifier. The leave-one-out cross validation was adopted for all comparisons.

1) Majority voting (MV) [98]. The linear SVM classifier was trained based on individual cell nuclei with labels being as same as set labels. Then each nucleus in the test set was predicted individually and the final set label was assigned by majority voting.

2) Feature statistics (STATS) [45]. Statistical features were extracted for all nuclear attributes. In the experiment, minimum, maximum, mean, standard deviation, skewness and kurtosis were

Table 3.1: CLASSIFICATION RESULTS BY DIFFERENT APPROACHES IN FOUR DIAGNOSTIC CHALLENGES (HAND-CRAFTED FEATURES)

Methods	FA vs. NG			FVPC vs. NG*			FNH vs. HCC*			DN vs. MM*		
	Accuracy	AUC	Cohen's kappa	Accuracy	AUC	Cohen's kappa	Accuracy	AUC	Cohen's kappa	Accuracy	AUC	Cohen's kappa
MV	69.64%	0.7895	0.3928	82.00%	0.8328	0.6400	83.72%	0.8891	0.6412	71.22%	0.7501	0.4207
BoW	67.85%	0.6901	0.3571	76.67%	0.8458	0.5304	84.89%	0.9163	0.6723	69.78%	0.7214	0.3936
STATS	73.21%	0.7844	0.4643	78.00%	0.7955	0.5557	88.37%	0.9253	0.7543	72.66%	0.7345	0.4520
EMD	71.43%	0.7934	0.4286	82.00%	**0.8750**	0.6434	86.05%	0.9140	0.6958	70.50%	0.7456	0.4065
mi-Graph	**78.57%**	0.8162	**0.5714**	84.00%	0.8445	0.6815	88.37%	0.9037	0.7592	71.94%	**0.7521**	0.4372
SetSVM	**78.57%**	**0.8291**	**0.5714**	**86.00%**	0.8687	**0.7173**	**90.69%**	**0.9299**	**0.8014**	**74.51%**	0.7460	**0.4893**

Table 3.2: CLASSIFICATION RESULTS BY DIFFERENT APPROACHES IN FOUR DIAGNOSTIC CHALLENGES (AUTOENCODER FEATURES)

Methods	FA vs. NG			FVPC vs. NG			FNH vs. HCC*			DN vs. MM*		
	Accuracy	AUC	Cohen's kappa	Accuracy	AUC	Cohen's kappa	Accuracy	AUC	Cohen's kappa	Accuracy	AUC	Cohen's kappa
MV	60.71%	0.6824	0.2143	84.00%	0.9075	0.6845	79.06%	0.8077	0.5577	66.19%	0.6907	0.3225
BoW	58.92%	0.6008	0.1786	84.66%	0.8717	0.6884	80.23%	0.8371	0.6050	64.75%	0.7061	0.2959
STATS	69.64%	0.7691	0.3929	84.00%	0.8539	0.6785	81.39%	0.8416	0.6186	70.50%	0.7119	0.4157
EMD	67.85%	0.6888	0.3571	86.00%	0.8782	0.7200	81.39%	0.8801	0.6109	68.35%	0.7152	0.3667
mi-Graph	73.21%	0.6952	0.4286	**90.00%**	0.8873	**0.7961**	83.72%	0.9005	0.6560	70.50%	0.7220	0.4151
SetSVM	**75.00%**	**0.8431**	**0.5000**	88.00%	**0.9091**	0.7541	**84.88%**	**0.9140**	**0.6818**	**72.42%**	**0.7463**	**0.4421**

Table 3.3: CLASSIFICATION RESULTS BY DIFFERENT APPROACHES IN FOUR DIAGNOSTIC CHALLENGES (TBM)

Methods	FA vs. NG			FVPC vs. NG*			FNH vs. HCC*			DN vs. MM*		
	Accuracy	AUC	Cohen's kappa	Accuracy	AUC	Cohen's kappa	Accuracy	AUC	Cohen's kappa	Accuracy	AUC	Cohen's kappa
MV	71.43%	0.7704	0.4286	84.00%	0.9010	0.6689	83.72%	0.8688	0.6412	80.58%	0.7904	0.6128
BoW	70.23%	0.7721	0.4048	82.00%	0.8831	0.6330	82.94%	0.8364	0.6277	77.59%	0.7770	0.5530
STATS	73.21%	0.7921	0.4643	88.00%	0.9075	0.7541	81.39%	0.8575	0.5943	78.41%	**0.8070**	0.5682
EMD	71.43%	0.7717	0.4286	86.00%	0.8929	0.7145	83.72%	0.8462	0.6488	76.26%	0.7767	0.5253
mi-Graph	73.21%	0.7856	0.4643	88.00%	0.9086	0.7588	83.72%	0.8554	0.6629	79.14%	0.7887	0.5837
SetSVM	**76.78%**	**0.8048**	**0.5357**	**90.00%**	**0.9188**	**0.8000**	**86.04%**	**0.8710**	**0.6958**	**82.01%**	0.7825	**0.6411**

*challenges where SetSVM provides statistically better performance

computed and concatenated into a single vector to represent the nuclei set.

3) Bag of words (BoW) [35]. The dictionary in BoW was built in the nuclei feature space with k-means using training data. The BoW model generated a histogram to represent each nuclei set by measuring the occurrences of each word in the nuclei set.

4) Earth mover distance (EMD) [39]. The earth mover distance was used to measure the distance between two sets based on an optimization process. The computed EMD for X_i and X_j is denoted as $d_{EMD}(X_i, X_j)$, which was applied in SVM by the RBF kernel $K(X_i, X_j) = e^{-\gamma d^2_{EMD}(X_i, X_j)}$.

5) mi-Graph [101]. Compared with standard multiple instance learning (MIL), mi-Graph was designed under relaxed MIL assumption. It constructs a graph for the nuclei set with nuclei being nodes and a predefined graph kernel is used in SVM to measure the similarity between two sets.

The performance of SetSVM was evaluated on the three datasets with four diagnostic challenges, namely, FA vs. NG, FVPC vs. NG, FNH vs. HCC and DN vs. MM. We calculated cancer detection accuracy to summarize the overall classification performance. In addition, the area under the receiver operating characteristic curve (AUC) and Cohens kappa coefficient were considered to evaluate the model.

Table 3.1, 3.2 and 3.3 contain the summary of classification results in each diagnostic challenge with hand-crafted features, autoencoder features and TBM respectively. The cancer detection performances were evaluated for SetSVM as well as five existing approaches mentioned above. As we can see from Table 3.1, 3.2 and 3.3, SetSVM provides statistically better classification accuracies in challenges marked with * (assessed by p values, $p < 0.05$) compared with five existing methods using different nuclear quantification approaches. The improvements are gained by optimizing both set representation and classifier decision boundary in one consistent cost function.

As far as computational complexity, our method only needs to compute the distance between each instance and each prototype to obtain the linear embeddings. Suppose we have m defined

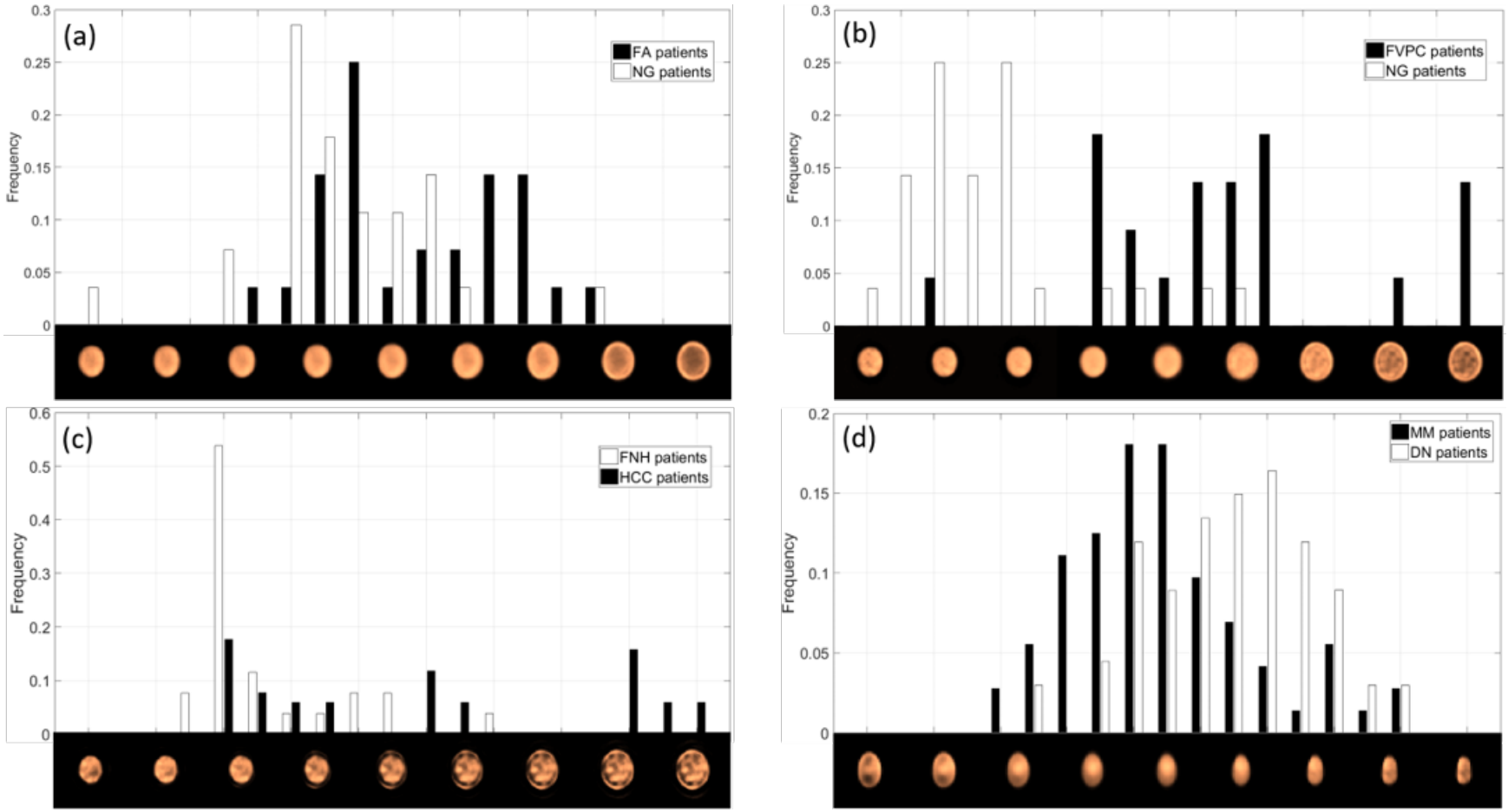

Figure 3.5: Distributions of discriminative nuclear patterns in four diagnostic challenges: follicular adenoma of the thyroid (FA) vs. nodular goiter (NG) (a), follicular variant of papillary thyroid carcinoma (FVPC) vs. nodular goiter (NG) (b), focal nodular hyperplasia (FNH) vs. malignant hepatocellular carcinoma (HCC) (c), and dysplastic nevi (DN) vs. malignant melanoma (MM) (d). For view purpose, the generated nuclei images beneath histogram bars are plotted in pseudo color.

prototypes and each set contains n instances, the computational complexity is $O(mn)$ to compute the instance-prototype distances for a set. However, in mi-Graph, the computational complexity is $O(n^2)$ to compute the instance-wise distances, which increases quadratically with the n.

3.4.2 Visualizing nuclear characteristics in different groups

SetSVM is able to provide visual exploration of the discriminant patterns for the nuclei set when using invertible nuclear quantifications (e.g. sparse autoencoder features). As mentioned in section 3.3.3, each set can be characterized by a feature vector in the nuclear feature space regarding the most important prototype p_o. We utilized the extracted set information in combination with $pLDA$ technique to visualize nuclear morphological differences between nuclei sets. For each diagnostic challenge, we projected all the data onto the discriminant direction V_{pLDA} with fixed $\epsilon = 0.001$ in 3.20 and the discriminant variations f_μ along V_{pLDA} is visualized by the inverse transformation in SSAE, as shown in the bottoms of Fig. 3.5 (a)-(d). The generated nuclei images beneath histogram bars are plotted in pseudo color. After projection, patient representations most similar to these nuclei images are counted in the corresponding bins. The height of bars in histograms indicates the frequency of corresponding morphometry patterns in each diagnostic challenge.

The histograms in Fig. 3.5 (a) suggests that nuclei in FA patients tend to have slightly bigger nuclei size and more chromatin concentrated around the nuclei membrane, while nuclei in NG patients are relatively small with more uniform chromatin distributed within the central region of the nucleus.

In comparison between FVPC and NG shown in Fig. 3.5 (b), the distribution histograms are more widely separated compared with FA vs. NG. A large proportion of FVPC patients have significantly bigger nuclei than NG patients. We computed the nuclei area of segmented nuclei images for all groups, shown in Table 3.4 and found that the average nuclear size in FVPC patients is 38% bigger than NG patients. Moreover, greater numbers of FVPC patients tend to have central clearing nuclei with peripheralization of chromatin compared with NG group.

Results for the challenge FNH vs. HCC are shown in Fig. 3.5 (c). It is clear that nuclei in the malignant group (HCC patients) are much bigger (34% bigger calculated from Table 3.4) than the benign group (FNH patients). In addition, HCC group tend to have more prominent chromatin mass condensed within the central region of nucleus, indicating more hyperchromatin inside nucleus than benign patients. In contrast, FNH patients have nuclei with relatively uniform chromatin distribution and little variations in size.

Fig. 3.5(d) shows the analysis result for comparison between DN and MM. The nuclei of DN patients are usually small and condensed without a nucleolus. However, MM patients tend to have enlarged nuclei (21% larger calculated from Table 3.4) with chromatin condensed around nuclei membrane and central region.

Table 3.4: Nuclear size (in pixels $\times 10^3$) in different groups

Groups	**FA**	**FVPC**	**NG**	**FNH**	**HCC**	**DN**	**MM**
Nuclear size	5.2 ± 1.5	6.9 ± 1.9	4.9 ± 1.3	1.6 ± 0.3	2.2 ± 0.8	0.5 ± 0.3	0.6 ± 0.2

3.5 Discussion

Nuclear structure, as observed under microscopy on routine staining, has long been prime interests of pathologists in cancer diagnosis. Models in CAD systems are required to solve the set classification problem and predict the presence or absence of cancer based on sets of nuclei. In this chapter, we described a novel set classification approach SetSVM in the application of nuclear morphometry-based cancer detection. SetSVM integrates set representation learning and classifier training in one unified cost function for better discriminativeness. The method is based on the idea of measuring set-set similarity by comparing matched nuclei obtained via a collection of prototypes. For better discriminative power, both decision boundary and prototypes are optimized to maximize the separation margin. SetSVM provides set representation to summarize characteristics of nuclear morphometry for any nuclei set. Representative information in the nu-

clei set can be visualized directly by feature decoding in the sparse stacked autoencoder (SSAE) in the discriminant subspace.

In nuclear-based cancer detection tasks, we compared our approach with five commonly used methods using thyroid, liver and melanoma datasets with 260 patients in total. All the methods took as input the same nuclear quantifications using hand-crafted features, autoencoder features and TBM. Experiment results show that the proposed SetSVM is able to provide state-of-art performances in almost all diagnostic challenges using different nuclear quantification approaches. One possible reason is that compared with SetSVM, other methods separate mapping function and classifier training, which may limit the model performance when the objectives of the two processes are inconsistent. In addition, the validation with three different types of features suggests that SetSVM is likely to provide superior performances independent of nuclear quantification approaches in CAD systems.

Beyond cancer detection, SetSVM enables visualization of discriminant nuclear patterns for the nuclei set. Combined with SSAE and $pLDA$, the modes of variations that are responsible for distinguishing two classes can be plotted in image space to discover the characteristic chromatin distribution within the nucleus. As far as we know, this is the first attempt to visually interpret such biological information at the patient level. In pathology, the spatial arrangement of chromatin is often associated with tumor progress. The compact, condensed dark stained heterochromatin inside nucleus reflects relatively low levels of transcriptional inactivity. Nuclear morphological features including chromatin clearing with peripheral margination of chromatin and nuclear enlargement have been documented as important diagnostic information in differentiating thyroid lesions. Enlargement in nuclear size and prominent nucleoli are often observed in malignant lesions and are used for cancer grading. Together with SSAE and $pLDA$, the proposed SetSVM provides a practical tool for direct visualization of representative nuclear patterns in patients from specific pathological lesions.

Although the proposed method can yield satisfying performance, it also has some limitations. First, SetSVM is built directly on set levels and only predicts labels for sets rather than instances.

Beyond label prediction for the patients, it would be interesting to extend the work to infer the probability of individual nuclei being a certain class. By doing so, we may be able to plot the heatmap and thus locate the tumor region in pathology images. Second, in this chapter, nuclear structure was quantified with unsupervised approaches (hand-crafted features, autoencoder features and TBM) and SetSVM only optimizes set representation and the classifier. There is still room to further improve the discriminativeness by introducing label information to nuclear quantification, leading to an end-end learning architecture.

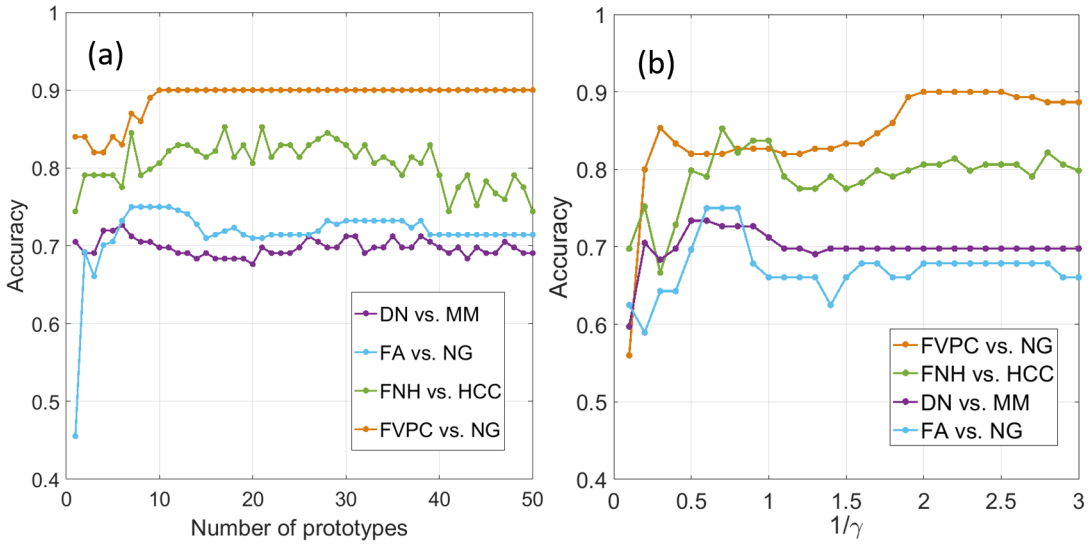

Figure 3.6: Performance analysis on sensitivity of parameters. (a) classification accuracy vs. number of prototypes; (b) classification accuracy vs. smoothing parameter. Here we provide the example analysis based on autoencoder features.

We note that in SetSVM two parameters: number of prototypes m and the smoothing parameter γ in mapping function are important to cancer detection performance. To test the sensitivity of each parameter, we fixed the other parameter as the optimal value and present the classification accuracy with respect to the change of m and γ in each diagnostic challenge, as shown in Fig. 3.6. The number of prototypes ranges from 1 to 50 and increasing m improves the performance. However, further increase of m would lead to suboptimal/unchanged classification accuracies. Lager number of prototypes m can generate higher dimensional set representations, where redundant information may degrade performance of the predictive model. Similar performance patterns can be observed when keeping m fixed and changing $\frac{1}{\gamma}$ from 0.1 to 3.0. When $\frac{1}{\gamma}$ is very

small, the mapping function $\Phi_{p_k}(X_i)$ becomes finding the nearest neighbor in X_i with respect to prototype p_k and thus Φ_{p_k} neglects characteristics of other instances. When $\frac{1}{\gamma}$ is too large, the nuclei characteristics summarized by prototypes within a set are 'obscured', losing discriminative features for classification.

Finally, it is worth noting that the proposed SetSVM is a quite general predictive model. It is a type of weakly supervised method in the sense that it takes as input sets of unlabeled feature vectors with only set-level labels available, alleviating the efforts to access detailed labels in many situations. In addition, SetSVM is able to visualize representative set information when combined with invertible features and can thus be extended to many pattern recognition tasks.

Chapter 4

Applications to General Pattern Recognition Problems

Predicting the class label for a set of instances is an important problem which is ubiquitous in many fields with data captured by various types of sensors. We believe that SetSVM is a quite general approach that can be applied in many situations beyond nuclei-based cancer detection in pathology. In this chapter, we present experimental evaluations showing that SetSVM can be used broadly in a variety of pattern recognition tasks, including mass classification in mammograms, classification of flow cytometry data as well as scene classification using natural images. Experimental results demonstrate that SetSVM can significantly improve classification accuracy compared to classic approaches in the corresponding fields.

4.1 Mass classification in mammograms

4.1.1 Introduction

In the US, breast cancer is currently the second most common cancer in women and may occur in men. Breast cancer starts from cells in the breast that grow out of control and usually form a tumor which is visible in x-ray images. A mammogram is a low-dose x-ray of the breast that

is used to look for signs of breast cancer, including small white spots, lumps or tumors called masses, and other suspicious areas. Mass is one of the major symptoms of the breast cancer and it is challenging to differentiate malignant masses from benign ones, due to large variations of mass shapes and image textures. For mass classification in mammograms, a benign mass is generally round or oval and has a well-defined boundary, whereas the malignant tumor is spiculated and has a blurry boundary. In the literature, shape-based features [32], texture-based features [63] and neighborhood intensity [56] are commonly used as metrics for mass classification. However, the performance of classification models using shape-based features, such as compactness, fractional concavity and fractal dimension, depends heavily on the accurate mass segmentation, which is difficult to obtain in many situations [56]. For the generality and simplicity, directly using intensity neighborhoods of pixels have been validated in mass classification [56], texture classification [82], object segmentation [11].

In image categorization, texton analysis often combines with various types of filter bank responses as an approach for texture modeling [55]. Texton analysis assumes that image texture is a collection of fundamental micro-structures, referred as textons, occurring repeatedly across the image [102]. Thus it is reasonable to expect that pixel representations form clusters in a certain feature space and that image textures can be modeled by the compact representation with a few textons. In this section, we used the intensity neighborhood ($m \times m$ pixel square, in $\mathbb{R}^{m^2 \times 1}$ feature space) to represent each image pixel. The k-means clustering algorithm is applied to find N_{dic} cluster centers, known as dictionary. Image pixels are then assigned to the corresponding nearest dictionaries with pre-defined distance metric. For image representation, the normalized probability density function is computed as a texture signature, which acts as the input of classifiers for pattern recognition. Texton analysis has shown satisfying performances in mass classification for differentiating cases from benign and malignant groups [56], [55], [8].

Here we consider the mass classification task as a set classification problem, where each image is a set consisting of various number of pixel representations without knowing the pixel-level annotations and the aim is to predict whether a mass region is from the benign or malignant

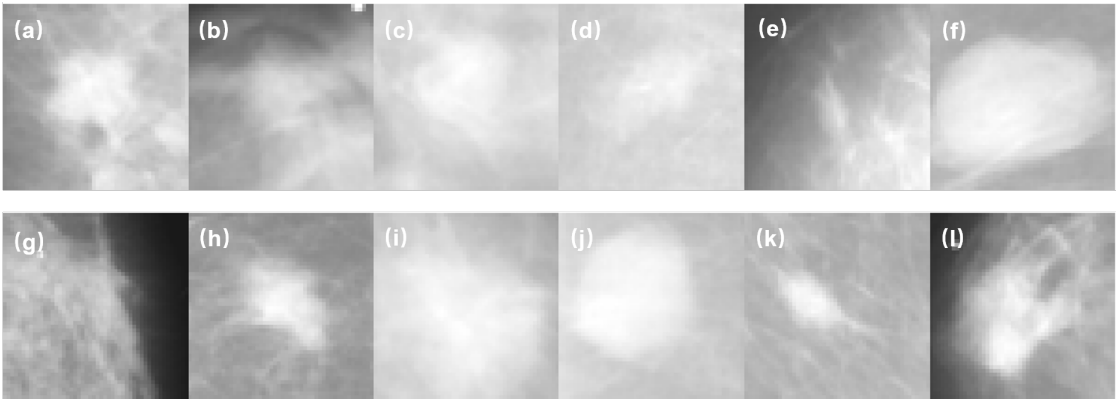

Figure 4.1: Sample mass regions from the DDSM dataset. (a)-(f) benign mass regions; (g)-(l) malignant mass regions.

Table 4.1: Mammogram dataset description

	1	2	3	4	5
Shape	11	31	32	37	3
Margin	40	16	18	14	26
Density	28	48	35	3	
Assessment	0	1	39	43	31
Subtlety	4	8	24	18	60

The first to the last columns show the ratings on a scale of 1-5 for mass properties of shape, margin, density, assessment as well as subtlety. The mass shape can display in round (1), oval (2), lobulated (3), irregular (4) and architecture distortion (5). The margin can be circumscribed (1), microlobulated (2), obscured (3), ill-defined (4)and spiculated (5). Assessment indicates the clinical severity of breast cancer and subtlety represents if the mass is obvious in mammograms where 1 is subtle and 5 is obvious [56].

group.

4.1.2 Dataset

In this experiment, we utilized the dataset from the Digital Database for Screening Mammography (DDSM) [1] [38], which is widely used in mammographic image analysis research. Mass regions were selected from two classes: benign cancer volumes and malignant cancer volumes. Images in the experiment were scanned on LUMISYS digitizer with 50μm/pixel at 12 bits/pixel.

[1] `http://marathon.csee.usf.edu/Mammography/Database.html`

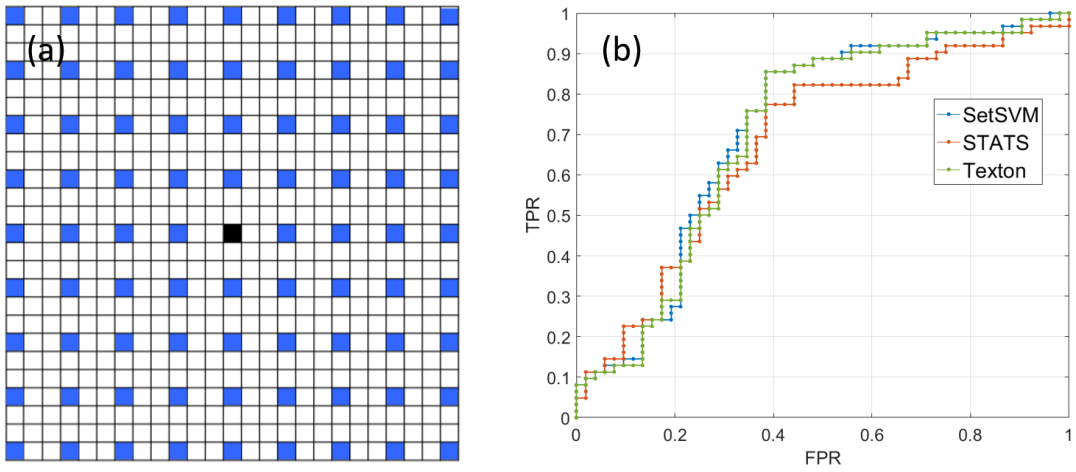

Figure 4.2: (a) Subsampling strategy in the 25×25 intensity neighborhood. Blue pixels are sampled for the black center pixel; (b) ROC curves for different classification approaches.

A total of 114 mass regions were included for study consisting of 52 benign masses and 62 malignant masses. Sample mass regions are shown in Figure 4.1. Since a small intensity neighborhood needs to be extracted for all pixels in mass regions, we down-sampled the images at the resolution of 200 μm/pixel to reduce computation complexity. The data statistics for mass properties are listed in Table 4.1.

4.1.3 Experiment results

To test the classification performace of different models, the 'leave-one-out' strategy was utilized in experiment comparisons. We opted the 25×25 image patch as the pixel neighborhood and applied sampling strategy (shown in Figure 4.2 (a) [56]) to reduce redundant information, resulting in a 9×9=81 dimensional feature vector to describe the local pattern of each pixel. In the experiment, we compared the performance of SetSVM with frequently used texton analysis as well as STATS 3 using the SVM classifier.

The experiment results are presented in Table 4.2 where classification accuracy, Cohen's kappa and AUC are listed for three methods. The accuracy was averaged on 20 individual executions. As we can see from Table 4.2, SetSVM outperforms the other two methods in classification accuracy and shows better agreement with the ground truth. In addition, receiver operating char-

Table 4.2: Classification accuracy comparison on benign vs. malignant (%)

Methods	Accuracy	Cohen's kappa	AUC
Texton analysis	71.93%	0.4430	0.7063
STATS	68.42%	0.3535	0.6743
SetSVM	**76.32%**	**0.5128**	**0.7146**

acteristics (ROC) curves are plotted in Figure 4.2 (b), showing the diagnostic ability of binary classifiers as one discrimination threshold changes. The standard Student's t-test confirms the statistically significant improvement by SetSVM with p-value less than 0.01.

4.2 Flow cytometry-based cancer detection

4.2.1 Introduction

Flow cytometry is a powerful tool in analyzing characteristics of particles flowing in the stream of fluid and has been widely applied in medical research and clinical practice [37]. Some examples of application fields include pathology, molecular biology, immunology and marine science as so on. Flow cytometry can yield multiple measurements for thousands of cells in a short time from light scatter and fluorescence emission signals. Forward scattered (FS) light is refracted by a cell and continues along in the light path, reflecting the cell size. Side scatter (SS) signal is collected at roughly 90 degrees from the original light path, reflecting cell internal complexity (i.e. granularity). Cell can be processed with fluorescent dyes or fluorescence-tagged antibodies, producing fluorescent light which reveals physiological and chemical properties of cells [7]. In a flow cytometer, a mixture of light signals are directed by optical filters and beam splitters and each signal is collected by the relevant detectors, generating electronic signals proportional to the signal that hit them. An illustrative diagram of the flow cytometer is shown in Figure 4.3.

In clinical practice, gating is a commonly applied method to select a subset of cells according to a small number of variables, which allows researcher to gather and display more information

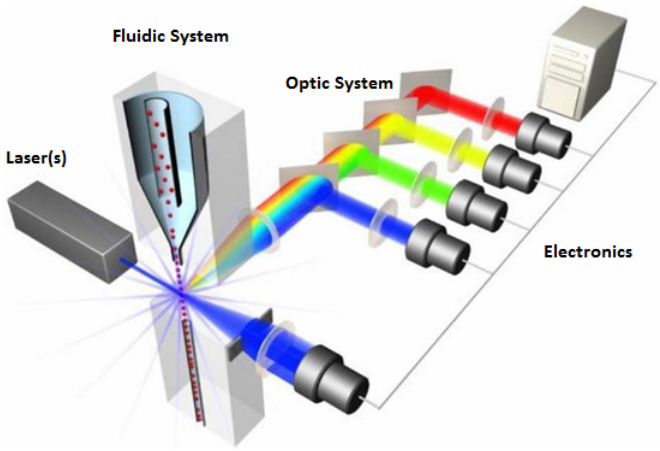

Figure 4.3: Overview of the flow cytometer. This image is taken from [73]

of cell subpopulation. A gate is usually a graphical boundary on a plot to isolate the desired cell subpopulation. The selected cell subpopulation are further analyzed based on remaining measurements and multiple gates can be combined in order to reliably identify a certain subtype of cells. Flow cytometry has been applied in the diagnosis of health conditions, especially blood diseases such as leukemia. Due to the large amount of available data and challenges in flow cytometry-based diagnosis, accurate and automated data analysis is a pressing need for clinical decision support [7].

Leukemia is a cancer of blood cells, which is often described as being acute or chronic. Acute myelocytic leukemia (AML) is characterized by the rapid growth of abnormal white blood cells, accounting for roughly 1.2% cancer deaths in the US [43]. Flow cytometry is a critical part of diagnosis of AML, allowing detection of aberrant protein expression profiles in monocytic cells [62]. In 2011, the DREAM 6/FlowCAP2 Molecular Classification of Acute Myeloid Leukemia (AML) Challenge[2] attracted a number of teams with the aim of developing machine learning algorithms to predict the condition for patients whose diagnosis was unknown to participants. In [59], feature expansion including multiplication and division between measurements was used and Fishers linear discriminant analysis (LDA) was then applied to reduce feature dimension.

[2]http://dreamchallenges.org/project/dream-6-flowcap2-molecular-classification-of-acute-myeloid-leukaemia-challenge/

The prediction score for any unknown patient was given based on the ℓ_1 regularized logistic regression model. In [7], six statistics regarding each measurement were aggregated as features. The Generalized Matrix Relevance Learning Vector Quantization (GMLVQ) was used for patient prediction.

In flow cytometry-based cancer detection, multiple measurements are provided for each cell and there are thousands of cells originating from a single patient who was diagnosed as normal or malignant. Instead of classifying each cell as positive or negative, the key is to consider the cell population as as whole and predict the patient condition. Therefore, the application of SetSVM to flow cytometry-based cancer detection is quite natural. In this section, we aim to apply the SetSVM model to predict the diagnosis of any test patient as AML-positive or healthy based on flow cytometry data.

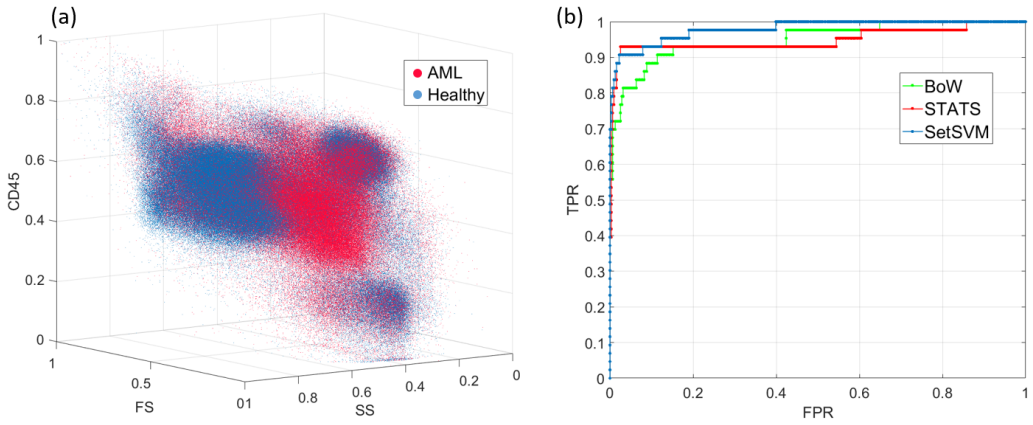

Figure 4.4: AML flow cytometry data visualized in FS, SS and CD45 space (a); ROC curves for different approaches on differentiating normal and AML patients (b)

4.2.2 Dataset description

We utilized the AML flow cytometry dataset in CSV format downloaded from the FlowRepository database [3], obtained from peripheral blood or bone marrow aspirates. The dataset consists of 359 subjects in total with 316 healthy patients and 43 AML-positive patients. Seven measurements for each cell were included in the experiment: forward scatter in linear scale (FS Lin),

[3] `https://flowrepository.org/id/FR-FCM-ZZYA`

Table 4.3: Classification accuracy comparison for differentiating normal from AML patients (%)

Methods	Accuracy	Cohen's kappa	AUC
BoW	95.82%	0.7772	0.9498
STATS	96.65%	0.8348	0.9493
SetSVM	**97.77%**	**0.8851**	**0.9668**

sideward scatter in logarithmic scale (SS Log), and five fluorescence intensities (FL1-FL5, IgG1-FITC, IgG1-PE, CD45-ECD, IgG1-PC5, IgG1-PC7) in logarithmic scales. The number of cells originating from each patient ranges from 6764 to 49370. In Figure 4.4(a), cells in the AML flow cytometry data are visualized in 3D space defined by FS Lin, SS Log and CD45-ECD.

4.2.3 Experiment results

In this experiment, 'leave-one-out' strategy was utilized to test the cancer detection performances of three methods: BoW, STATS and SetSVM. Cell measurements were normalized between 0 to 1 at the same scale. In [7], statistics including mean, standard deviation, skewness, kurtosis, median and interquartile range were used along with a non-linear classifier to produce 100% classification accuracy. For STATS in the our experiment, we used the exact same six statistics as in [7] to represent cell characteristics. Table 4.3 shows the classification results for methods evaluated by average accuracy, Cohen's kappa and AUC. We should note that the line of chance is 88.02% in this task and we found that the misclassified cases are mostly from AML group in three methods due to the unbalanced data. The ROC curves for methods are plotted in Figure 4.4(b).

4.3 Natural scene classification

4.3.1 Introduction

Much of the recent progress in computer vision has been made by designing robust image features or classifiers for the task of scene classification. Scene classification aims to automatically categorize the environment in a given image as belonging to one of a set of scene classes, like mountain, beach, desert, etc. Scene classification is useful in place recognition, image retrieval, multimedia direct marketing and so on. Although great effort has been made, it is still a challenging task due to many factors to be considered such as illumination, object scale and position [54].

In the literature, many features are based on low-level image properties (*e.g.* SIFT [58], filterbank responses [29]) and have achieved success in image scene classification. The popular bag of words (BoW) model often combines with the extracted low-level features to represent an image as a collection of local features [105]. Based on the framework of BoW, many variants are developed for scene classification. In [99], multi-resolution representation was involved in BoW model, where local image features were extracted from multi-resolution images. In [52], spatial pyramid matching (SPM) was proposed to incorporate global geometric correspondence, where an image is divided into small cells and concatenate the histogram of cells to the histogram of original images using low-level features. Low-level features have been proven effective in many scene classification tasks, however, its performance may be limited as the visual tasks become challenging due to the small local area size [54].

Superpixel segmentation is an approach to aggregate a collection of pixels with shared features such as color, illumination, spatial location and can provide higher level information compared with pixel-level responses. In addition, superpixel-level feature analysis is more efficient than pixel-level features since the number of superpixels is much less than that of pixels and thus it help reduce computation complexity. Features are extracted from each superpixel region to describe the image characteristics including color, texture, shape. In [92], covnolutional neural

network (CNN) was applied to extract features automatically from superpixels for segmenting and classifying epithelial and stromal regions. In [81], local binary patterns (LBP) features were used in very high resolution image (VHR) classification. Recently, the CNN model has been widely applied in image pattern recognition tasks and has been the state-of-art approach in scene classification [51].

In this section, we view the scene classification as a set classification problem where an image consists of a set of pixel-level or superpixel-level feature descriptors and determine where the image was taken.

4.3.2 Dataset

In this section, we aim to classify images into scene categories sky vs. desert. The image data is a subset of the SUN [4] database, which consists of 168 images taken from scene sky and 202 images taken from scene desert. Images in both categories are with various sizes and a few samples are shown in Figure. 4.5 (a) and (b).

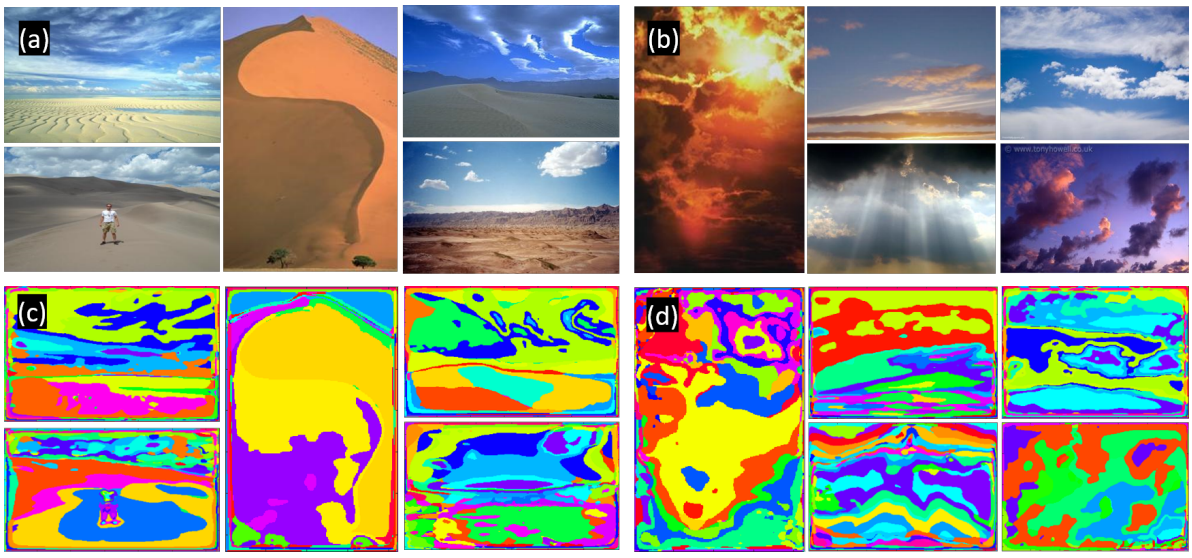

Figure 4.5: Sample images and corresponding output maps of visual words. (a) sample desert images; (b) sample sky images; (c) maps of visual words for (a); (d) maps of visual words for (b).

[4] http://groups.csail.mit.edu/vision/SUN/

4.3.3 Method

Pixel-level feature extraction

To extract pixel-level features, we convert the digital image from RGB color space to LAB color space using the method presented in [25]. The characteristics of LAB color make it suitable for color image feature extraction. LAB color space is designed to approximate human vision, where luminance is represented on the L axis, perpendicular on a pile of 'ab' plane with uniformly distributed colors from green to red along 'a' axis and from blue to yellow along the 'b' axis. The multi-scale filter bank (33 filters in total, as shown in Figure 4.6), including Gaussian filters, Laplacian of Gaussian (LoG) filters and first derivative of Gaussian filters at different scales, is applied to L, A, B channels separately of the image to generate pixel-level filter responses.

Create visual words for BoW model

The dictionary of visual words can be constructed using *k*-means clustering algorithm. Instead of using all the pixel-level responses, we randomly select α pixels from each image. If there are T training images, the obtained filter responses matrix would be $\alpha T \times N$, where N is the number of filter responses per pixel. Each pixel in an image is mapped to its closest word in the dictionary measured by the standard Euclidean distance, producing a map of visual words where each pixel is assigned the index of its closest visual word. Several maps of visual words are visualized in Figure 4.5 (c) and (d).

Superpixel-level feature extraction

Images are partitioned into a number of superpixels, which should accurately adhere to the object boundaries and should be computation efficient. Simple linear iterative clustering (SLIC) algorithm [1] is a commonly used method for superpixel generation by performing a local clustering of pixels in the 5-D space defined by the L, A, B values in the CIELAB color space as well as the *x, y* coordinates.

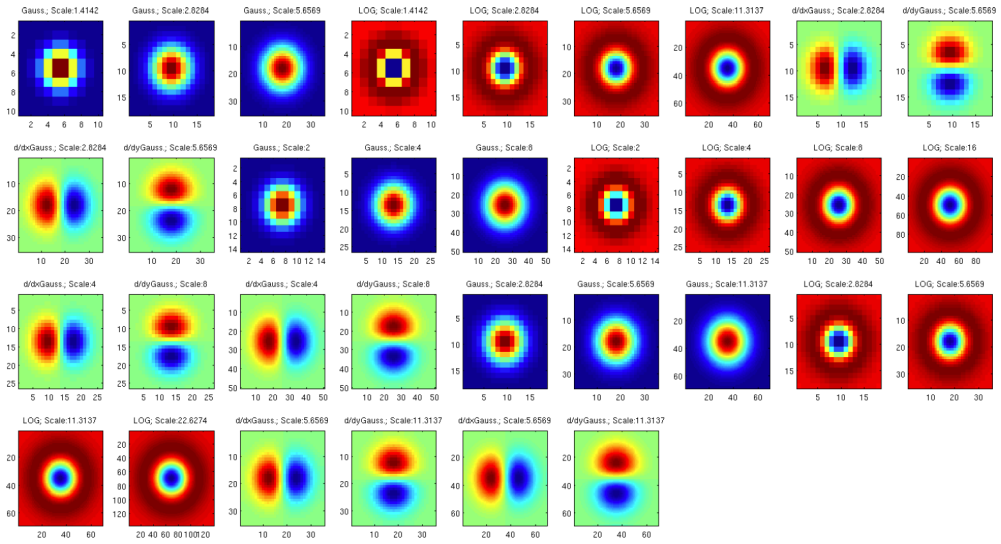

Figure 4.6: The applied multi-scale filter bank to generate pixel responses.

For an image with N pixels and K desired superpixels, SLIC algorithm begins with initializing K evenly distributed cluster centers with grid interval $S = \sqrt{N/K}$ and moves cluster centers to the lowest gradient position in the 3×3 neighborhood. The K cluster centers are denoted as $C_k = [l_k, a_k, b_k, x_k, y_k]^T$, with $k = [1, K]$. For each cluster center, assign the best matching pixels in the $2S \times 2S$ image neighborhood using the distance measure as follows:

$$\begin{aligned} d_{lab} &= \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2} \\ d_{xy} &= \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \\ D &= d_{lab} + \frac{m}{S} d_{xy} \end{aligned} \tag{4.1}$$

where D is the distance between any pixel i and cluster center C_k by summing up distances in LAB color space and coordinate space. m is a parameter to control the compactness of the superpixel within the value range $[1, 20]$.

After image pixels are assigned to their corresponding closest cluster centers, the average 5-dimensional vector over the pixels within the same superpixel is computed as the new center. Such process is repeated iteratively until the changes of cluster centers are sufficiently small. The SLIC algorithm is computation efficient with $O(N)$ complex. A few examples of superpixel segmentation are shown in Figure 4.7, where each image contains various numbers of superpixels.

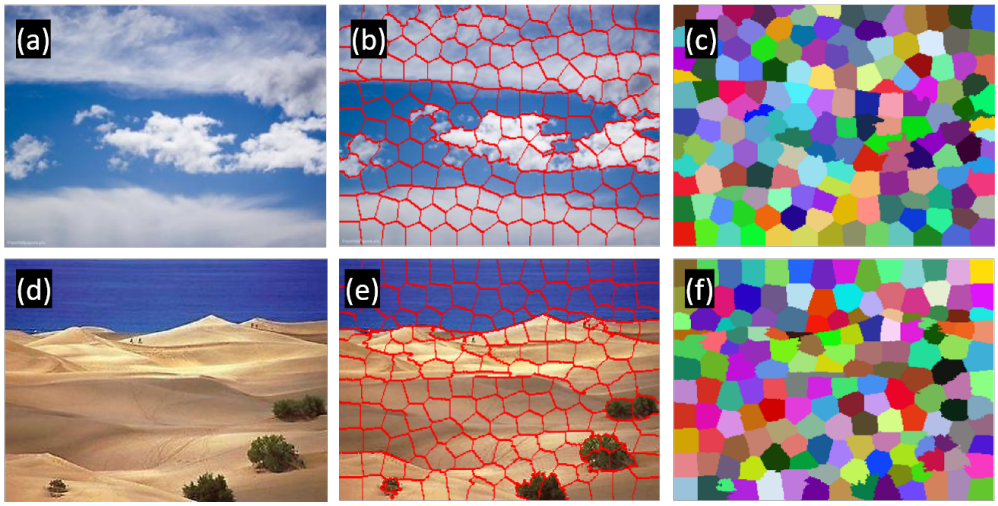

Figure 4.7: Example images and their SLIC superpixel segmentation. Input sky image (a) and desert image (d); boundary overlay (b) and (e); pixels within the same superpixel are assigned random colors (c) and (f).

Haralick features [36] are then applied to quantify texture characteristics of superpixel regions in three image channels, generating a set of feature vectors for each image. Haralick's texture features are computed based on $N_g \times N_g$ gray-level co-occurrence matrix (GLCM) with N_g being the number of gray levels of the image. Each element $[i, j]$ in GLCM indicates the probability of a pixel with value i adjacent to another pixel with value j. Haralick features are often calculated from GLCMs generating from each of the four directions: horizontal, vertical, left and right diagonals. Haralick features normally are 13 types of statistics extracted from the GLCM: angular second moment (ASM), contrast, correlation, variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, information measure of correlation 1 and information measure of correlation 2. Refer to Murphy's lab [5] for details about Haralick feature calculation.

4.3.4 Experiment results

At pixel-level, 33 filters were convolved with each of the three image channels separately in LAB color space , generating 99 responses per pixel. We selected $\alpha = 150$ pixels randomly from each

[5] `http://murphylab.web.cmu.edu/publications/boland/boland_node26.html`

of training images to construct $K = 200$ visual words in the BoW model. All images were represented as 200-dimensional feature vectors showing the occurrences of the visual words. At superpixel-level, each image was firstly partitioned into around 100 superpixels using SLIC, each of which is considered as an instance of the image. With GLCMs calculated at 6 gray-scale levels, 468 Haralick features were extracted from three image channels for any superpixel region. For both pixel-level features and superpixel-level features, the standard principle component analysis (PCA) technique was then applied to the entire feature set and the top 10 feature directions that captured more than 95% variations were retained to describe each instance.

Table 4.4: Classification results on sky vs. desert with pixel-level features

Methods	Accuracy	Cohen's kappa	AUC
BoW	86.22%	0.7179	0.9092
STATS	89.18%	0.7421	0.9376
CNN	**92.97%**	**0.8583**	**0.9751**
SetSVM	91.35%	0.8257	0.9577

Table 4.5: Classification results on sky vs. desert with superpixel-level features

Methods	Accuracy	Cohen's kappa	AUC
BoW	84.05%	0.6756	0.9104
STATS	90.27%	0.8041	0.9680
CNN	92.97%	0.8583	0.9751
SetSVM	**93.51%**	**0.8693**	**0.9790**

In the experiment, we compared the scene classification performances of SetSVM with BoW model as well as STATS with both pixel-level features and superpixel-level features, as shown in Table 4.4 and Table 4.5. In addition, we also tested the classification performance of the CNN model. Due to the small scale of the test dataset, we utilized a one-layer CNN, where 15 filters with size 5×5 were used to produce the feature map. The one-layer CNN model consists of a convolution layer, a max-pooling layer, a RELU activation layer, and a softmax output layer. Experiment results showed that SetSVM outperformed BoW and STATS using pixel-level features. In Table 4.4, the CNN model performed the best with classification accuracy of 92.97%. In addition, we found that compared with pixel-level features extracted within small local regions, higher-level information embedded in superpixels may help improve classification performances for STATS and SetSVM. As shown in Table 4.5, SetSVM achieved the highest

classification accuracy of 93.51%. The receiver operating characteristics (ROC) curves for the four methods are plotted in Figure 4.8 (a) and (b).

As far as computation complexity, BoW, STATS and SetSVM took 140s, 34s and 173s respectively using superpixel-level features, while the computation time was 1600s, 529s and 2056s respectively with pixel-level features. The one-layer CNN model took 3606s in the experiment. All comparisons were performed on a machine with Intel(R) Core(TM) i7 4810MQ 2.8GHz CPU and 12 Gb RAM.

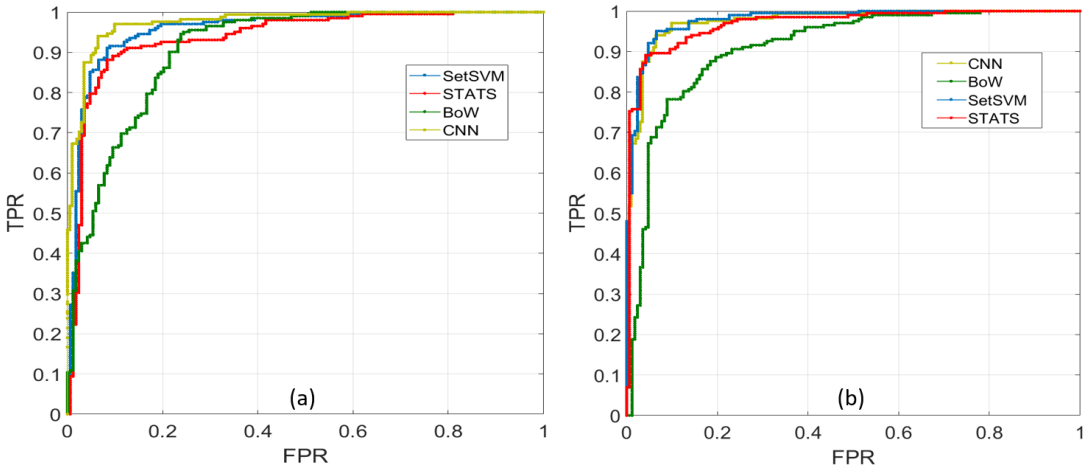

Figure 4.8: ROC curves for different methods using pixel-level features (a) and superpixel-level features (b). Note that the same ROC curves for CNN are included in both (a) and (b) for comparison.

Chapter 5

Conclusions

Nuclear morphology is an important indicator of cellular processes and plays a significant role in decision making for disease diagnosis. By imaging large numbers of slides automatically at high resolution, digital pathology has the potential to become a useful tool in pathology practice, facilitating pathologists decisions, and overall benefiting the patient. In this thesis, we propose novel nuclei detection, nuclei segmentation and cancer detection algorithms in the image analysis pipeline to maximize the overall amount of information extracted from nuclei images.

Segmentation is a critical prerequisite in systems for quantitative analysis of nuclear morphology. In Chapter 2, we introduced an unsupervised method, which is called multi-scale edge selection in polar space (MESPS), to efficiently locate and extract nuclei without pre-training. Potential nuclei can be located automatically by measuring the matching degrees between local image regions and a set of filters. In the proposed method, nuclei segmentation problem becomes searching for the shortest path between two nodes in an undirected graph. Compared with supervised segmentation that requires a large number of labeled samples for model training, our method is able to adapt to different types of nuclei datasets and provides similar or even better performance. Qualitative and quantitative analysis showed that the method is automatic and accurate when segmenting nuclei from variously stained images with noisy background and has the potential to be used in clinic settings.

Even though MESPS achieved satisfying segmentation results in experimental validations, we admit that the amount of datasets we used is far from enough compared with the large variety of nuclei appearances displayed in pathology images. During the last few decades, many segmentation solutions have been proposed in the literature, unfortunately, nuclei segmentation still remains an unsolved problem. One major challenge is the lack of high-quality public benchmark dataset and the lack of recognized segmentation metrics to evaluate and compare different approaches. Many methods are evaluated on their own datasets with various metrics, making it extremely difficult to recognize a new solution as a state-of-art approach in the field. For algorithm development in the future, transfer learning is one potential direction for nuclei segmentation since it combines the merits of both supervised learning and unsupervised learning. Transfer learning is able to 'borrow' knowledge learned from training samples (source domain) and perform the same task on the unseen datasets coming from other sources (target domain), largely reducing the amount of work for manual annotations.

We formulated the nuclei-based cancer detection task as the set classification problem and proposed a novel predictive model SetSVM in Chapter 3. Different from approaches trying to build instance level classifier and then adopt a certain voting strategy for set label prediction, SetSVM considers a set of instances as a whole without any assumption, aiming to train a classifier directly at set level. Compared with existing set classification approaches which consist of two optimization steps with inconsistent objective functions, SetSVM unifies set representation learning with classifier training in one step to maximize model's overall discrimination ability. The method solves the set classification problem by jointly optimizing the mapping function and the SVM decision boundary in a maximum soft margin cost function. We showed that a better performance is possible by introducing discriminant information from the classifier to the mapping function. Using multiple nuclear quantification approaches, experiment results showed that SetSVM provides significant improvements compared with state-of-art approaches in cancer detection tasks including thyroid cancer, liver cancer and melanoma. In addition, we showed that SetSVM enables visual interpretation of discriminative nuclear characteristics representing the

nuclei set. These features make SetSVM a potentially practical tool in building accurate and interpretable CAD systems for cancer detection.

Set classification is relatively a new topic compared with single instance classification. We believe set classification is a quite general strategy to many problems and provide experimental validations for SetSVM in several pattern recognition tasks in Chapter 5, including mass classification in mammograms, cancer detection based on flow cytometry data as well as natural scene classification. Experiment results demonstrated that SetSVM can provide significant improvements compared with state-of-art approaches in corresponding fields. In addition, its performance seems independent of types of data be captured and instance features. Note that predicting the class label for a set of unlabeled instances is an important and ubiquitous problem with many applications. SetSVM can thus be applied to other similar tasks such as quantitative assessment of drug effects and video sequence classification.

However, we should mention that the average execution time of SetSVM tested on the MATLAB platform is longer than that of BoW and STATS approaches, but less than the mi-Graph and CNN models. Better optimization methods and faster programming languages (*e.g.* C, C++) are needed to reduce the computation complexity, especially for large scale datasets. In this thesis, we implemented SetSVM based on linear support vector machine classifier to transfer discriminative information to set representation learning. We should note that the idea of 'end-end training' in SetSVM can be extended to any non-linear and differentiable classifier for better class separation ability, *e.g.* neural network classifier for multi-classification applications.

Finally, we mention the analogy between the set classification problem and multiple instance learning (MIL). Since MIL was defined with strict assumption by Dietterich et al. [21] and there is no consensus on whether the relaxed version of MIL problem belongs within the MIL scope, we use the term 'set classification problem' instead. The superior performance of set classification has already been validated in [3] using seven databases from different domains of knowledge. Our experiment results confirmed the conclusion in [3] that discriminative classifier should be based on global information from the whole set to predict the set label. Recently, the

combination of MIL and the deep learning model is attracting more and more attention in vision related applications [88], [79]. However, these ideas are based on the strict MIL assumption which may be problematic in nuclei-based cancer detection tasks. In the future, set classification with relaxed MIL assumption provides a possible direction for designing powerful deep learning models in numerous pattern recognition problems.

Bibliography

[1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. 4.3.3

[2] Yousef Al-Kofahi, Wiem Lassoued, William Lee, and Badrinath Roysam. Improved automatic detection and segmentation of cell nuclei in histopathology images. *IEEE Transactions on Biomedical Engineering*, 57(4):841–852, 2010. 1.2.1

[3] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013. 1.2.2, 5

[4] Vahid Anari, Parvin Mahzouni, and Rasoul Amirfattahi. Computer-aided detection of proliferative cells and mitosis index in immunohistichemically images of meningioma. In *Machine Vision and Image Processing (MVIP), 2010 6th Iranian*, pages 1–5. IEEE, 2010. 1.2.1

[5] Chamidu Atupelage, Hiroshi Nagahashi, Fumikazu Kimura, Masahiro Yamaguchi, Abe Tokiya, Akinori Hashiguchi, and Michiie Sakamoto. Computational hepatocellular carcinoma tumor grading based on cell nuclei classification. *Journal of Medical Imaging*, 1(3): 034501–034501, 2014. 1.2.2

[6] William Beaver, David Kosman, Gary Tedeschi, Ethan Bier, William McGinnis, and Yoav Freund. Segmentation of nuclei in confocal image stacks using performance based thresholding. In *Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*, pages 53–56. IEEE, 2007. 1.2.1

[7] Michael Biehl, Kerstin Bunte, and Petra Schneider. Analysis of flow cytometry data by matrix relevance learning vector quantization. *PLoS One*, 8(3):e59401, 2013. 4.2.1, 4.2.1, 4.2.3

[8] Anna Bosch, Xavier Munoz, Arnau Oliver, and Joan Marti. Modeling and classifying breast tissue density in mammograms. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1552–1558. IEEE, 2006. 4.1.1

[9] Hakan Cevikalp and Bill Triggs. Face recognition based on image sets. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2567–2573. IEEE, 2010. 1.2.2

[10] Tony F Chan and Luminita A Vese. Active contours without edges. *IEEE Transactions on image processing*, 10(2):266–277, 2001. 1.2.1

[11] Cheng Chen, John A Ozolek, Wei Wang, and Gustavo K Rohde. A general system for automatic biomedical image segmentation using intensity neighborhoods. *Journal of Biomedical Imaging*, 2011:8, 2011. 4.1.1

[12] Cheng Chen, Wei Wang, John A Ozolek, and Gustavo K Rohde. A flexible and robust approach for segmenting cell nuclei from 2d microscopy images using supervised learning and template matching. *Cytometry Part A*, 83(5):495–507, 2013. 1.2.1, 2.3.2, 3.3.2

[13] Kin-Hoe Chow, Rachel E Factor, and Katharine S Ullman. The nuclear envelope environment and its cancer connections. *Nature reviews. Cancer*, 12(3):196, 2012. 1.1

[14] Dan C Cireşan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber. Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 411–418. Springer, 2013. 1.2.1

[15] Luís Pedro Coelho, Aabid Shariff, and Robert F Murphy. Nuclear segmentation in microscope cell images: a hand-segmented dataset and comparison of algorithms. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on*, pages

518–521. IEEE, 2009. 2.3.3

[16] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619, 2002. 1.2.1

[17] Eric Cosatto, Matt Miller, Hans Peter Graf, and John S Meyer. Grading nuclear pleomorphism on histological micrographs. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008. 1.2.1

[18] Celine Denais and Jan Lammerding. Nuclear mechanics in cancer. In *Cancer Biology and the Nuclear Envelope*, pages 435–470. Springer, 2014. 1.1

[19] Atam P Dhawan and Louis Arata. Segmentation of medical images through competitive learning. *Computer Methods and Programs in Biomedicine*, 40(3):203–215, 1993. 1.2.1

[20] Atam P Dhawan and Anne Sim. Segmentation of images of skin lesions using color and texture information of surface pigmentation. *Computerized Medical Imaging and Graphics*, 16(3):163–177, 1992. 1.2.1

[21] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1):31–71, 1997. 5

[22] Edsger W Dijkstra. A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271, 1959. 2.2.2

[23] Scott Doyle, Shannon Agner, Anant Madabhushi, Michael Feldman, and John Tomaszewski. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pages 496–499. IEEE, 2008. 1.2

[24] M Murat Dundar, Sunil Badve, Gokhan Bilgin, Vikas Raykar, Rohit Jain, Olcay Sertel, and Metin N Gurcan. Computerized classification of intraductal breast lesions using

histopathological images. *IEEE Transactions on Biomedical Engineering*, 58(7):1977–1984, 2011. 1.2.2

[25] Mark D Fairchild. *Color appearance models*. John Wiley & Sons, 2013. 4.3.3

[26] Hussain Fatakdawala, Jun Xu, Ajay Basavanhally, Gyan Bhanot, Shridar Ganesan, Michael Feldman, John E Tomaszewski, and Anant Madabhushi. Expectation–maximization-driven geodesic active contour with overlap resolution (emagacor): Application to lymphocyte segmentation on breast cancer histopathology. *IEEE Transactions on Biomedical Engineering*, 57(7):1676–1689, 2010. 1.2.1

[27] Pawel Filipczuk, Thomas Fevens, Adam Krzyzak, and Roman Monczak. Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. *IEEE transactions on medical imaging*, 32(12):2169–2178, 2013. 1.2.2

[28] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2.2.2

[29] William T Freeman, Edward H Adelson, et al. The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, 13(9):891–906, 1991. 4.3.1

[30] Thomas Gärtner, Peter A Flach, Adam Kowalczyk, and Alexander J Smola. Multi-instance kernels. In *ICML*, volume 2, pages 179–186, 2002. 1.2.2

[31] Alexandru Grumezescu and Anton Ficai. *Nanostructures for Cancer Therapy*. Elsevier, 2017. 1.1

[32] Denise Guliato, Rangaraj M Rangayyan, Juliano D Carvalho, and Sérgio A Santiago. Polygonal modeling of contours of breast tumors with the preservation of spicules. *IEEE Transactions on Biomedical Engineering*, 55(1):14–20, 2008. 4.1.1

[33] Metin N Gurcan, Tony Pan, Hiro Shimada, and Joel Saltz. Image analysis for neuroblastoma classification: Segmentation of cell nuclei. In *Engineering in Medicine and Biology*

Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE, pages 4844–4847. IEEE, 2006. 1.2.1

[34] Metin N Gurcan, Laura E Boucheron, Ali Can, Anant Madabhushi, Nasir M Rajpoot, and Bulent Yener. Histopathological image analysis: A review. *IEEE reviews in biomedical engineering*, 2:147–171, 2009. 1.2, 1.2, 1.2.2

[35] Ju Han, Yunfu Wang, Weidong Cai, Alexander Borowsky, Bahram Parvin, and Hang Chang. Integrative analysis of cellular morphometric context reveals clinically relevant signatures in lower grade glioma. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 72–80. Springer, 2016. 1.2.2, 3.4.1

[36] Robert M Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979. 4.3.3

[37] S Sakira Hassan, Pekka Ruusuvuori, Leena Latonen, and Heikki Huttunen. Flow cytometry-based classification in cancer research: a view on feature selection. *Cancer informatics*, 14(Suppl 5):75, 2015. 4.2.1

[38] Michael Heath, Kevin Bowyer, Daniel Kopans, Richard Moore, and W Philip Kegelmeyer. The digital database for screening mammography. In *Proceedings of the 5th international workshop on digital mammography*, pages 212–218. Medical Physics Publishing, 2000. 4.1.2

[39] Hu Huang, Akif Burak Tosun, Jia Guo, Cheng Chen, Wei Wang, John A Ozolek, and Gustavo K Rohde. Cancer diagnosis by nuclear morphometry using spatial information. *Pattern recognition letters*, 42:115–121, 2014. 1.2.2, 3.4.1

[40] Humayun Irshad, Antoine Veillard, Ludovic Roux, and Daniel Racoceanu. Methods for nuclei detection, segmentation, and classification in digital histopathology: a reviewcurrent status and future potential. *IEEE reviews in biomedical engineering*, 7:97–114, 2014. 1.1, 1.2, 1.2.1

[41] Sezgin M Ismail, Angela B Colclough, John S Dinnen, Douglas Eakins, DM Evans, Ernest

Gradwell, Jerry P O'sullivan, Joan M Summerell, and Robert G Newcombe. Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia. *BMJ*, 298(6675):707–710, 1989. 1.1

[42] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016. 2.1

[43] Ahmedin Jemal, Andrea Thomas, Taylor Murray, and Michael Thun. Cancer statistics, 2002. *CA: a cancer journal for clinicians*, 52(1):23–47, 2002. 4.2.1

[44] Chanho Jung and Changick Kim. Impact of the accuracy of automatic segmentation of cell nuclei clusters on classification of thyroid follicular lesions. *Cytometry Part A*, 85(8): 709–718, 2014. 1.2.1

[45] Melih Kandemir and Fred A Hamprecht. Computer-aided diagnosis from weak supervision: A benchmarking study. *Computerized Medical Imaging and Graphics*, 42:44–50, 2015. 3.4.1

[46] Adnan Mujahid Khan, Nasir Rajpoot, Darren Treanor, and Derek Magee. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Transactions on Biomedical Engineering*, 61(6):1729–1738, 2014. 1.2

[47] Parmeshwar Khurd, Leo Grady, Ali Kamen, Summer Gibbs-Strauss, Elizabeth M Genega, and John V Frangioni. Network cycle features: Application to computer-aided gleason grading of prostate cancer histopathological images. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pages 1632–1636. IEEE, 2011. 1.2.1

[48] Teuvo Kohonen. Improved versions of learning vector quantization. In *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*, pages 545–550. IEEE, 1990. 3.2.4

[49] Hui Kong, Metin Gurcan, and Kamel Belkacem-Boussaid. Partitioning histopathological images: an integrated framework for supervised color-texture segmentation and cell

splitting. *IEEE transactions on medical imaging*, 30(9):1661–1677, 2011. 1.2.1

[50] Sonal Kothari, John H Phan, Todd H Stokes, and May D Wang. Pathology imaging informatics for quantitative analysis of whole-slide images. *Journal of the American Medical Informatics Association*, 20(6):1099–1108, 2013. 1.2

[51] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 4.3.1

[52] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2169–2178. IEEE, 2006. 4.3.1

[53] Chunming Li, Chenyang Xu, Changfeng Gui, and Martin D Fox. Distance regularized level set evolution and its application to image segmentation. *IEEE Transactions on image processing*, 19(12):3243–3254, 2010. 2.3.2

[54] Li-Jia Li, Hao Su, Yongwhan Lim, and Li Fei-Fei. Objects as attributes for scene classification. In *European Conference on Computer Vision*, pages 57–69. Springer, 2010. 4.3.1

[55] Xi-Zhao Li, Simon Williams, Gobert Lee, and Min Deng. Computer-aided mammography classification of malignant mass regions and normal regions based on novel texton features. In *Control Automation Robotics & Vision (ICARCV), 2012 12th International Conference on*, pages 1431–1436. IEEE, 2012. 4.1.1

[56] Yanfeng Li, Houjin Chen, Gustavo Kunde Rohde, Chang Yao, and Lin Cheng. Texton analysis for mass classification in mammograms. *Pattern Recognition Letters*, 52:87–93, 2015. 4.1.1, 4.1, 4.1.3

[57] Gang Lin, Umesh Adiga, Kathy Olson, John F Guzowski, Carol A Barnes, and Badrinath Roysam. A hybrid 3d watershed algorithm incorporating gradient cues and object models

for automatic segmentation of nuclei in confocal image stacks. *Cytometry Part A*, 56(1): 23–36, 2003. 1.2.1

[58] David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999. 4.3.1

[59] Tapio Manninen, Heikki Huttunen, Pekka Ruusuvuori, and Matti Nykter. Leukemia prediction using sparse logistic regression. *PloS one*, 8(8):e72932, 2013. 4.2.1

[60] Ke Zhi Mao, Peng Zhao, and Puay-Hoon Tan. Supervised learning-based cell image segmentation for p53 immunohistochemistry. *IEEE Transactions on Biomedical Engineering*, 53(6):1153–1163, 2006. 1.2.1

[61] Hela Masmoudi, Stephen M Hewitt, Nicholas Petrick, Kyle J Myers, and Marios A Gavrielides. Automated quantitative assessment of her-2/neu immunohistochemical expression in breast cancer. *IEEE transactions on medical imaging*, 28(6):916–925, 2009. 1.2.1

[62] Sergio Matarraz, Julia Almeida, Juan Flores-Montero, Quentin Lécrevisse, Valentina Guerri, Antonio López, Susana Bárrena, Vincent HJ Van Der Velden, Jeroen G Te Marvelde, Jacques JM Van Dongen, et al. Introduction to the diagnosis and classification of monocytic-lineage leukemias by flow cytometry. *Cytometry Part B: Clinical Cytometry*, 92(3):218–227, 2017. 4.2.1

[63] Loris Nanni, Sheryl Brahnam, and Alessandra Lumini. A very high performing system to discriminate tissues in mammograms as benign and malignant. *Expert Systems with Applications*, 39(2):1968–1971, 2012. 4.1.1

[64] Ahmed M Nazif and Martin D Levine. Low level image segmentation: An expert system. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):555–577, 1984. 1.2.1

[65] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Trans-*

actions on systems, man, and cybernetics, 9(1):62–66, 1979. 1.2.1

[66] Mohammad Peikari, Mehrdad J Gangeh, Judit Zubovits, Gina Clarke, and Anne L Martel. Triaging diagnostically relevant regions from pathology whole slides of breast cancer: A texture based approach. *IEEE transactions on medical imaging*, 35(1):307–315, 2016. 1.2.2

[67] Hady Ahmady Phoulady, Mu Zhou, Dmitry B Goldgof, Lawrence O Hall, and Peter R Mouton. Automatic quantification and classification of cervical cancer via adaptive nucleus shape modeling. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 2658–2662. IEEE, 2016. 1.2.2

[68] Marina E Plissiti and Christophoros Nikou. Overlapping cell nuclei segmentation using a spatially adaptive active physical model. *IEEE Transactions on Image Processing*, 21 (11):4568–4580, 2012. 1.2.1

[69] Xin Qi, Fuyong Xing, David J Foran, and Lin Yang. Robust segmentation of overlapping cells in histopathology specimens using parallel seed detection and repulsive level set. *IEEE Transactions on Biomedical Engineering*, 59(3):754–765, 2012. 1.2.1

[70] Gwenolé Quellec, Mathieu Lamard, Michel Cozic, Gouenou Coatrieux, and Guy Cazuguel. Multiple-instance learning for anomaly detection in digital mammography. *IEEE transactions on medical imaging*, 35(7):1604–1614, 2016. 1.2.2

[71] Francesco Raimondo, Marios A Gavrielides, Georgia Karayannopoulou, Kleoniki Lyroudia, Ioannis Pitas, and Ioannis Kostopoulos. Automated evaluation of her-2/neu status in breast tissue from fluorescent in situ hybridization images. *IEEE Transactions on Image Processing*, 14(9):1288–1299, 2005. 1.2.1

[72] Gustavo K Rohde, John A Ozolek, Anil V Parwani, Liron Pantanowitz, et al. Carnegie mellon university bioimaging day 2014: Challenges and opportunities in digital pathology. *Journal of pathology informatics*, 5(1):32, 2014. 1.1, 1.2.1

[73] Tom Rowley. Flow cytometry-a survey and the basics. *Columbia University*, 2013. (doc-

ument), 4.3

[74] Arnout C Ruifrok, Dennis A Johnston, et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299, 2001. (document), 2.3, 2.3.2

[75] Oliver Schmitt and Maria Hasse. Morphological multiscale decomposition of connected regions with emphasis on cell clusters. *Computer Vision and Image Understanding*, 113 (2):188–201, 2009. 1.2.1

[76] M Alper Selver, Aykut Kocaoğlu, Güleser K Demir, Hatice Doğan, Oğuz Dicle, and Cüneyt Güzeliş. Patient oriented and robust automatic liver segmentation for pre-evaluation of liver transplantation. *Computers in Biology and Medicine*, 38(7):765–784, 2008. 2.3.3

[77] Ajit Singh, Demetri Terzopoulos, and Dmitry B Goldgof. *Deformable models in medical image analysis*. IEEE Computer Society Press, 1998. 1.2.1

[78] Korsuk Sirinukunwattana, Adnan M Khan, and Nasir M Rajpoot. Cell words: Modelling the visual appearance of cells in histopathology images. *Computerized Medical Imaging and Graphics*, 42:16–24, 2015. 1.2.1

[79] Miao Sun, Tony X Han, Ming-Chang Liu, and Ahmad Khodayari-Rostamabad. Multiple instance learning convolutional neural networks for object recognition. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3270–3275. IEEE, 2016. 5

[80] Philippe Thevenaz, Ricard Delgado-Gonzalo, and Michael Unser. The ovuscule. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):382–393, 2011. 2.3.2

[81] John E Vargas, Priscila TM Saito, Alexandre X Falcao, Pedro J de Rezende, and Jefersson A dos Santos. Superpixel-based interactive classification of very high resolution images. In *Graphics, Patterns and Images (SIBGRAPI), 2014 27th SIBGRAPI Conference on*, pages 173–179. IEEE, 2014. 4.3.1

[82] Manik Varma and Andrew Zisserman. Texture classification: Are filter banks necessary? In *Computer vision and pattern recognition, 2003. Proceedings. 2003 IEEE computer society conference on*, volume 2, pages II–691. IEEE, 2003. 4.1.1

[83] Mitko Veta, Josien PW Pluim, Paul J Van Diest, and Max A Viergever. Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411, 2014. 1.2

[84] Christian Wallraven, Barbara Caputo, and Arnulf BA Graf. Recognition with local features: the kernel recipe. In *ICCV*, volume 3, pages 257–264, 2003. 1.2.2, 3.2.1

[85] Jun Wang and Jean-Daniel Zucker. Solving multiple-instance problem: A lazy learning approach. 2000. 1.2.2

[86] Wei Wang, Yilin Mo, John A Ozolek, and Gustavo K Rohde. Penalized fisher discriminant analysis and its application to image-based morphometry. *Pattern recognition letters*, 32 (15):2128–2135, 2011. 3.3.3

[87] Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013. 3.3.3

[88] Jiajun Wu, Yinan Yu, Chang Huang, and Kai Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3460–3469, 2015. 5

[89] Xuqing Wu, Mojgan Amrikachi, and Shishir K Shah. Embedding topic discovery in conditional random fields model for segmenting nuclei using multispectral data. *IEEE Transactions on Biomedical Engineering*, 59(6):1539–1549, 2012. 1.2.1

[90] Fuyong Xing and Lin Yang. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE reviews in biomedical engineering*, 9:234–263, 2016. 1.2.1

[91] Fuyong Xing, Yuanpu Xie, and Lin Yang. An automatic learning-based framework for

robust nucleus segmentation. *IEEE transactions on medical imaging*, 35(2):550–566, 2016. 2.1

[92] Jun Xu, Xiaofei Luo, Guanhao Wang, Hannah Gilmore, and Anant Madabhushi. A deep convolutional neural network for segmenting and classifying epithelial and stromal regions in histopathological images. *Neurocomputing*, 191:214–223, 2016. 4.3.1

[93] Jun Xu, Lei Xiang, Qingshan Liu, Hannah Gilmore, Jianzhong Wu, Jinghai Tang, and Anant Madabhushi. Stacked sparse autoencoder (ssae) for nuclei detection on breast cancer histopathology images. *IEEE transactions on medical imaging*, 35(1):119–130, 2016. 1.2.1, 2.1, 3.3.3

[94] Yan Xu, Jun-Yan Zhu, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical image analysis*, 18(3):591–604, 2014. 1.2.2

[95] Anthony Yezzi, Satyanad Kichenassamy, Arun Kumar, Peter Olver, and Allen Tannenbaum. A geometric snake model for segmentation of medical imagery. *IEEE Transactions on medical imaging*, 16(2):199–209, 1997. 1.2.1

[96] Kun-Hsing Yu, Ce Zhang, Gerald J Berry, Russ B Altman, Christopher Ré, Daniel L Rubin, and Michael Snyder. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature Communications*, 7, 2016. 1.2, 1.2.2, 3.1

[97] Bo Zhang, Christophe Zimmer, and J-C Olivo-Marin. Tracking fluorescent cells with coupled geometric active contours. In *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*, pages 476–479. IEEE, 2004. 1.2.1

[98] Xiaofan Zhang, Fuyong Xing, Hai Su, Lin Yang, and Shaoting Zhang. High-throughput histopathological image analysis via robust cell segmentation and hashing. *Medical image analysis*, 26(1):306–315, 2015. 1.2.2, 3.4.1

[99] Li Zhou, Zongtan Zhou, and Dewen Hu. Scene classification using a multi-resolution

bag-of-features model. *Pattern Recognition*, 46(1):424–433, 2013. 4.3.1

[100] Xiaobo Zhou and Stephen TC Wong. Informatics challenges of high-throughput microscopy. *IEEE Signal Processing Magazine*, 23(3):63–72, 2006. 1.2

[101] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-iid samples. In *Proceedings of the 26th annual international conference on machine learning*, pages 1249–1256. ACM, 2009. 1.2.2, 3.4.1

[102] Song-Chun Zhu, Cheng-En Guo, Yizhou Wang, and Zijian Xu. What are textons? *International Journal of Computer Vision*, 62(1):121–143, 2005. 4.1.1

[103] Christophe Zimmer, Elisabeth Labruyere, Vannary Meas-Yedid, Nancy Guillen, and J-C Olivo-Marin. Segmentation and tracking of migrating cells in videomicroscopy with parametric active contours: A tool for cell-based drug testing. *IEEE transactions on medical imaging*, 21(10):1212–1221, 2002. 1.2.1

[104] Daniele Zink, Andrew H Fischer, and Jeffrey A Nickerson. Nuclear structure in cancer cells. *Nature reviews. Cancer*, 4(9):677, 2004. (document), 1.1, 1.1

[105] Jinyi Zou, Wei Li, Chen Chen, and Qian Du. Scene classification using local and global features with collaborative representation fusion. *Information Sciences*, 348:209–226, 2016. 4.3.1