LARGE-SCALE ANALYSIS, MANAGEMENT, AND RETRIEVAL OF BIOLOGICAL AND MEDICAL IMAGES

A Dissertation

presented to

the Faculty of the Graduate School

University of Missouri – Columbia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

JING HAN

Dr. Chi-Ren Shyu, Dissertation Supervisor

MAY 2016

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

LARGE-SCALE ANALYSIS, MANAGEMENT, AND RETRIEVAL OF BIOLOGICAL AND MEDICAL IMAGES

presented by Jing Han,

a candidate for the degree of doctor of philosophy,

and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Chi-Ren Shyu

Dr. Gerald Arthur

Dr. Prasad Calyam

Dr. Guilherme DeSouza

DEDICATION

To my Dad: I did it! Our dream finally came true!

To my Mom: You are my rock!

To my Husband: Without your unconditional love and support, I wouldn't have made it this far.

To Cooper and Trevor: 妈妈爱你们!

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest gratitude to my academic advisor, Dr. Chi-Ren Shyu. He has consistently challenged me to be a better student and researcher, provided invaluable suggestions and ideas when obstacles were encountered, and been a constant source of encouragement and motivation. Additionally, I would like to thank the members of my committee, Drs. Gerald Arthur, Prasad Calyam, and Guilherme DeSouza, for their feedback and suggestions during this process, which have made this dissertation stronger.

This work would also not be possible without the expertise and support of our domain experts. A special thanks goes out to Dr. Gerald Arthur and Dr. Dmitriy Shin for sharing their expertise in pathology informatics; to Dr. Karen Edison and Dr. Jonathan Dyer for letting me participate in their department sessions and shadow around their clinics; to Dr. Dmitry Korkin for introducing me into structural bioinformatics.

I would not have reached to this milestone without the friendship and help from my lab members at the interdisciplinary Data Analytics and Search (iDAS) Lab. I would specifically like to thank Drs. Jaturon Harnsomburana, Jason Green, Hongfei Cao, and Nan Zhao for their ideas, suggestions, assistance, and support throughout the development of this work. Mike Phinney, Matt Spencer, and Devin Petersohn also helped me to tackle so many issues while I was finding my ways to the Big Data world. A final thanks goes to my funding sources, including grant from the National Science Foundation (#DBI-1053024).

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LARGE-SCALE ANALYSIS, MANAGEMENT, AND RETRIEVAL OF BIOLOGICAL AND MEDICAL IMAGES

JING HAN

Dr. Chi-Ren Shyu, Dissertation Supervisor

ABSTRACT

Biomedical image data have been growing quickly in volume, speed, and complexity, and there is an increasing reliance on the analysis of these data. Biomedical scientists are in need of efficient and accurate analyses of large-scale imaging data, as well as innovative retrieval methods for visually similar imagery across a large-scale data collection to assist complex study in biological and medical applications. Moreover, biomedical images rely on increased resolution to capture subtle phenotypes of diseases, but this poses a challenge for clinicians to sift through haystacks of visual cues to make informative diagnoses. To tackle these challenges, we developed computational methods for large-scale analysis of biological and medical imaging data using simulated annealing to improve the quality of image feature extraction. Furthermore, we designed a Big Data infrastructure for the large-scale image analysis and retrieval of digital pathology images and conducted a longitudinal study of clinician's usage patterns of an image database management system (MDID) to shed light on the potential adoption of new informatics tools. This research also resulted in image analysis, management, and retrieval applications relevant to dermatology, radiology, pathology, life sciences, and palynology disciplines. These tools provide the potential to answer research questions that would not be answerable without our novel innovations that take advantage of Big Data technologies.

CHAPTER ONE

INTRODUCTION

In the realm of scientific studies where observers rely heavily on their visual perception of the subjects of interest to make discoveries and conclusions, the nature of the world at large is presented with and thus perceived by their visually salient features, such as colors, shapes, textures, spatial placement/displacement, as well as the changes in these aforementioned aspects. The imagery of a subject is captured either directly by eyes (the biological sensors) or indirectly via digitally captured media (the digital sensors) such as images and videos. Subsequently, the content of imagery can be examined, manipulated, compared, and archived for further study. It is argued that perception, especially vision, is not purely a passive receipt of external signals, but rather a rational process that "requires intelligent problem-solving based on knowledge" [1].

Many disciplines, including biology and medicine, require accurate characterization of visual patterns and rich content in imagery in order to make discoveries and diagnoses. In the biology disciplines, the study of biological components and their morphology can lead to discoveries of distortions or changes that contribute to functional abnormalities. For example, the dynamics of mitochondria in *Drosophila* segmental nerves have essential impacts on the functions of neurons. Their defects are strongly associated with many neurodegenerative diseases [2]. In medical disciplines (e.g. radiology, dermatology, and pathology) the study of patient images (e.g. CT's of lungs,

images of skin lesions, or microscopic slides of tumor biopsy) includes the examination of gross appearance as well as detailed morphology of individual organelles, cells, and tissue. The diagnosis is made with extensive reasoning that requires years of professional training and experience in each medical specialty. For example, a hematologist examines a glass slide of patient's bone marrow biopsy under a microscope, identifies frequent occurrences of centroblasts, a type of white blood cell, based on the perception of cell morphology (including but not limited to enlarged cell size, moderate to scant cytoplasm, and non-cleaved nucleus), and subsequently make a diagnosis of the patient with a specific type of blood disorder, such as diffuse large B-cell lymphoma (DLBCL).

Even with intensive training and accumulative experience, the judgment and characterization of visual patterns and rich content residing in the images can still be subjective, implicit, indirect, and oftentimes inconsistent between observers and among different observations from the same professional. This is where computer vision, image processing algorithms, and machine learning techniques play their roles in performing more sensitive and accurate measurements and analyses. They provide useful means for researchers to detect subtle changes or easily overlooked patterns by human observers. Additionally, computational methods also enable high-throughput analysis of large-scale datasets where pure human examination may reach its physical limits. The severity of such phenomena only increases with the volume, variety, veracity, and velocity of data being generated electronically by modern technologies. Researchers are in need of not only accurate and quick analyses of large-scale data but also advanced and smart retrieval of visually similar imagery across

large collections of data to assist complex studies in their domains. An automatic retrieval system would need to extract useful and representative features from the images themselves, learn from domain experts' advanced reasoning to analyze image content, make numerical measurements and comparisons across the entire collection, and eventually return a limited set of images that best qualify researchers' request of informational image query. Such systems need to be efficient, accurate, adaptable, and able to handle large-scale data processing.

1.1 Problem Statement

Due to the reliance on imaging to capture visual pattern and rich content of subjects of interest and the ever-growing scale of data collection, methods and applications are needed to assist large-scale computational processing and analysis of image data. In this dissertation, methods are developed to address several common issues in multiple medical and biological imaging domains, such as radiology, dermatology, pathology, life sciences, and palynology. With limited adjustment, these methods and techniques can be adapted to other domains that share similar reliance on imagery and visual content analyses.

A reliable image analysis system requires the following abilities: (1) meaningful objects of interest can be identified; (2) selected visual features are indeed sufficient to characterize the various appearances in the image dataset; and (3) feature values are properly extracted and accurately represent the true patterns residing in the images. These requirements sound fairly basic and intuitive, but are difficult to ensure at all times in practice. A significant amount of this difficulty lies in adjusting and adapting all of the various parameters used

in the object segmentation and feature extraction algorithms. This problem is common to most of the applications in image processing, analysis, and retrieval and is therefore a very valuable issue to be addressed.

The overwhelming amount of image data being captured daily in various research domains demands the development of matching computational skills. This is especially true in biomedical imaging research, as it is virtually impossible to acquire enough time, money, and personnel to manually annotate all images, with consistency and efficiency both ensured. Therefore, computer-assisted annotation and analysis tools would be valuable to achieve the required consistency and efficiency in image analysis and retrieval, making use of the content of images themselves, machine learning techniques, and modern applications, such as databases, data indexing, web servers, and the Big Data ecosystem.

To fulfill the usefulness of image analysis and retrieval systems and to achieve successful adoption in real-world practices, the usability of such systems needs to be seriously considered. Unfortunately, this aspect is often overlooked in the scientific domains, especially the life sciences and medical professions, as compared to social sciences.

1.2 Dissertation Organization

The rest of this dissertation is organized into the following chapters. Chapter 2 gives a broad introduction of basic concepts and common practices involved in imaging informatics, particularly for medical and biological domains. These items are revisited multiple times in subsequent chapters and thus deserve

a formal statement at the beginning. Chapter 3 presents research works related to image management, especially web-based applications for biological and clinical images. System usability and user behavioral patterns are also studied and presented in this chapter to demonstrate the importance of developing user-friendly systems for a successful adoption and positive contribution to domain-specific imaging informatics needs. The topics proceed in chapter 4 to the extraction of visual content in images from various research domains. Special attention was made to showcase our automatic parameter tuning approach that improves the quality of image segmentation and feature extraction. Although it was applied originally for radiology images, it has a wide range of potential applications in the general fields of imaging informatics. Chapter 5 demonstrates our work on content-based image retrieval in the medical and biological domains. First of all, multiple perceptual categories that radiologists utilize during examination and diagnosis are studied and formulated as computational modules targeting specific disease patterns. Using these modules, we developed an entropy-based multi-module content-based image retrieval (CBIR) system for HRCT images of lung. Furthermore, we constructed a relational database of Neotropical pollen and spore grains and their visual content. As one of the aspects of image content, grain morphology was emphasized in this CBIR system for microscopic images of fossil grains. Attention is then turned to the domain of digital pathology in Chapter 6 where the challenges of analyzing various visual patterns in microscopic virtual slides were addressed with methods in image processing and realized with modern techniques in Big Data ecosystem. The

dissertation then ends with conclusions and discussions of continuous works in the future in Chapter 7.

In summary, this dissertation's contributions are both, (1) developing applications in large-scale medical and biological image analysis and retrieval across multiple research domains, and (2) evaluating efficiency and usability of developed systems in the real-world practices.

CHAPTER TWO

BASIC CONCEPTS AND COMMON PRACTICES IN IMAGING INFORMATICS

Generally speaking, several aspects of imaging informatics are repeatedly referenced across multiple chapters in this dissertation. It is therefore worth the effort to introduce basic concepts, definitions and common practices that this dissertation relies upon. This chapter is by no means a comprehensive introduction of image processing, but rather a brief coverage of a select collection of topics under the umbrella of imaging informatics for biological and medical domains.

2.1 Basic Concepts

2.1.1 Digital Images

A digital image can be considered as a 2-dimensional matrix of pixels or a function of values, $f(x, y)$, where the values at each spatial location, depicted by its x- and y-coordinates, are scalars with one or multiple channels defined by chosen color systems. To simplify, the pixel value of a monochrome image at location (x, y) ranges between [*Lmin*, *Lmax*], which is called grayscale. Commonly, *Lmin* is set to be 0 representing pure black and *Lmax* is set to be l representing pure white. The value of l is determined based on data structure that stores grayscale values. For example, an 8-bit data structure has the maximal grayscale value of 255 for pure white, while a 16-bit data structure represents pure white with a value of 65535. For color images, a single pixel is formed with multiple values from predefined color channels. Each channel represents the intensity from corresponding color component. For example, the most common

color system is the Red, Green, and Blue (RGB) system. A pixel with value (R, G, B) = (255, 255, 0) is of color yellow as a combination of pure red and pure green.

2.1.2 Image Processing, Image Analysis, and Computer Vision

As suggested by Gonzalez and Woods, there is no clear boundary in the definitions of image processing, image analysis, and computer vision [3]. Computer vision, as a sub-discipline of artificial intelligence (AI), is designed to ultimately emulate human vision by collecting visual input, making inference, and performing actions as if it were a human being. Image analysis, or image understanding, is somewhere in between image processing and computer vision.

2.1.3 Regions and Objects of Interest

The definition of a region of interest (ROI) varies among disciplines. In image processing, we generally define it as a group of pixels that are clustered together, forming a region bearing meaningful concepts in specific domains. There can be multiple ROIs in an image with various arrangements, spatially and/or hierarchically. For example, a CT image of lung usually contains a general ROI of body (separated from air around the body region), which is further recognized to include two lung ROIs (in most cases) that themselves may contain detailed ROIs representing anatomically meaningful structures, such as trachea, bronchi, bronchioles, pulmonary arteries and veins, etc.

In the domain of image processing, the term “object of interest” is also frequently used to describe the domain specific regions that usually are the targeted subject of said research. For example, centroblasts (one type of white blood cells) are the objects of interest in the biopsied nodule tissue, which itself is

also a ROI on the virtual slide. The usage of regions and objects of interest is interchangeable most of the times in literature references and practices.

2.1.4 Visual Features and Image Content

As explained in Chapter 1, the visual perception of an image requires past knowledge, learning, memory, etc. In biomedical imaging informatics, this requirement translates loosely to the perception of visual characteristics of objects of interest, in terms of their color intensities, morphological patterns, special placement and/or displacement, textures, etc. Visual features of regions and/or objects of interest numerically represent these characteristics. For example, we use a set of measurements introduced in [4] to represent the textural patterns inside images. As another example, the morphology of an object can be described using several visual features such as size (the number of pixels that compose an object of interest), perimeter, aspect ratio, elongation, form factor, etc. For a complete list of commonly used visual features, please refer to Table 2.1 and Table 2.2.

The content of an image is essentially the information that an image contains or otherwise is perceived by observers (human or computer). For example, an image of a beach is typically recognized based on its key content i.e. some blue skies, ocean, sand, and possibly some palm trees. In this section, we are not trying to formally define what the image content is but rather informally acknowledge that in imaging informatics, we are not analyzing any annotated text about an image. Instead, analyzing, understanding, and utilizing the image content itself is the ultimate goal.

2.2 Common Practices

Even though we develop customized image processing and analysis pipelines for each study, as can be seen in the following chapters, there are a few fundamental components that occur repeatedly in various applications. In this section, we selectively discuss some of the common practices before introducing specific case studies.

2.2.1 Image Segmentation

Image segmentation, sometimes referred to as object segmentation, is a collection of digital image operations that partition an image into disjoint "segments" as regions and/or objects that are homogeneous inside and more dissimilar between individual regions [5]. Autonomous and robust image segmentation would generally lead to more successful recognition of individual objects and thus lead to more accurate measurements. At the same time, it is considered to be the most difficult step in the image processing pipeline. Partly because it is one of the early stages in the image processing pipeline, right after preprocessing steps such as image enhancement, and restoration [3], image segmentation largely determines the performance of the end results of image processing [6].

The classic image segmentation algorithms can be grouped into two categories: (1) those that divide regions based on the sudden discontinuity (edge-based) and (2) those that build up the coverage of regions by including neighboring pixels that are similar in characteristics (region-based). There is a wide collection of image segmentation methods. The most frequently used

methods are edge detection, thresholding, watersheding, and morphological operations (erosion, dilation, opening, closing, top-hat, black hat, etc.). Although these algorithms oftentimes operate on monochrome images, their extensions and variations are also developed to allow the processing of multi-channel color images [7].

Thresholding is one of the most popular algorithms for image segmentation. The algorithm assumes that objects are formed with similar pixel intensities within them in an image. It utilized gray-level histograms constructed from raw image/region pixels. A single threshold is selected to separate objects of interest in the image (global thresholding) or in local regions (adaptive thresholding). The threshold values can be empirically determined based on the developer's understanding of the nature of image collections in hand; may be computed from pre-segmented images as training samples; or may be determined solely on the pixel information contained in the image. In most cases, the last approach (unsupervised thresholding) is utilized since it is not always easy to have training samples, and empirically determined values could easily be biased with limited *a priori* knowledge. In our experience, the Otsu threshold [8] often works well in simple scenarios where there is a relatively clear separation observed in image pixel histogram. However, its effectiveness of separating foreground objects from background pixels diminishes as the content of the image gets more complex and the intensities of similar objects do not necessarily share similar ranges. That is where adaptive thresholding [9] plays its role in segmenting objects based on local surroundings.

Based on the assumption that regions tend to have sudden and local discontinuity of pixel values on their boundaries, another set of image segmentation algorithms operate on images by identifying such abrupt changes. Edge detection algorithms work well when there is a clear boundary present that separates regions from background. In practice, this assumption does not always hold, as boundaries in natural scene image as well as domain-specific images tend to be blurred and noisy to certain degrees. To overcome such ambiguity, different kernels can be utilized to model a *ramp* of pixel changes instead of a *step* edge. The steeper the ramp slope is, the clearer the edge appears. The common edge detection kernels include Sobel, Prewitt, and Laplacian operators [3].

Aside from these aforementioned common image segmentation algorithms, there is another set of operations that is also useful in finding the correct division between objects and background. They are called morphological image processing algorithms. This is a big family of algorithms stemming from the fundamental operations of erosion and dilation. Simply put, erosion is an operation that "shrinks" brighter objects by applying a small image (so-called structuring element, *SE*) over the target image, anchoring on each pixel, and making decisions whether certain criteria are met in order to mark a pixel to be kept as an object pixel or treated as a background pixel. The dilation operation is considered as the dual of erosion. This duality is also true for other advanced morphological operations that are often defined in duels, for example morphological opening (the dilation of an eroded image) is a duel of morphological closing (the erosion of an dilated image) [3, 10]. Once defined in

grayscale images, these algorithms can be extended to handle images with arbitrary dimensions and channels.

Furthermore, grayscale morphological reconstruction is developed on top of classic grayscale image morphological operations, yet it is considered to be one of the geodesic operations [11, 12]. The reconstruction is essentially the recovering of objects in the target image (so-called mask image) that are "marked" by another image (so-called marker image) and discarding those unmarked ones. However, when the target image and/or marker image are of grayscale pixel values, the operation will be more complex than simply picking marked connected components in mask image. First of all, a structuring element (*SE*) needs to be defined in the form of grayscale values. Then the geodesic distance is determined based on the connectivity in image grid system. The grayscale reconstruction by *dilation* of a grayscale mask image, g, by a grayscale marker image, f, is the iterative geodesic dilation of f with respect to g until stability is reached [11, 10]. The morphological reconstruction of a grayscale image by *erosion* can be easily defined by duality as mentioned in the previous paragraph. Similarly, there is a family of such geodesic operations on grayscale images for reconstruction. Their usage is best demonstrated by finding appropriate markers for Watersheding algorithms. Therefore, a robust grayscale image reconstruction algorithm can be beneficial to successful image segmentation due to its ability to separate overlapping objects while preserving as much of the original morphology as possible. We will discuss the details of this family of operations and their utilization in cell segmentation in Chapter 6.

2.2.2 Feature Extraction

Feature extraction involves the selection of a set of visual features that best describe the visual patterns of objects of interest. As a common subsequent step following image segmentation, the description of a segmented object is the quantitative representation that a computer can process. As introduced in section 2.1.4, there are two ways to describe object appearance: (1) its external characteristics (boundaries) and (2) its internal characteristics (pixels within). Each representation is materialized by numeric measurements that are carefully designed to truly extract the underlying features of objects. Regional features best describe the internal characteristics of an object, such as color and texture, while boundary representation focuses on the external shape, such as smoothness and convexity. It is, however, difficult to determine the suitable set of features that truly represent the complete and relevant information that humans usually find in the image and sometimes non-detectable by human eyes but salient to computer processing. The intuitive selection of features makes an attempt to reproduce the same patterns that a human would observe. For example, the basic statistics of pixel values (raw image) are considered low-level features. However, it is usually the set of high-level features that are derived from low-level features that play the determinative roles in object recognition, machine learning, and information retrieval [13]. With computational techniques, we are able to extract groups of high-level visual features in the following general types.

In the domain of digital image analysis, **color** is one of the most common types of feature extracted from images. Their values are invariant to translation and rotation and are usually insensitive to scale and occlusion changes [14].

Moreover, color is a rich representation of objects in a scene and humans can differentiate thousands of color shades, as opposed to relatively few grayscale shades. Colors are represented in the realm of digital image processing with different special color systems. The most common color system is the Red, Green, and Blue (RGB) color system in which each pixel comprises three values signifying the intensities captured by each color channel. This model seems to be non-intuitive for humans, but it originated from experimental evidence of the biology of how human eyes capture and divide colors. There are over 6 to 7 million cones in the human eye to sense color by measuring these three primary colors [3]. In practice, there are alternative color modeling systems that are proven to be as useful as, if not superior than, the RGB system. The most popular color systems are: HSV/I (Hue, Saturation, Value/Intensity), CMY (Cyan, Magenta, Yellow), and CIE L*a*b (Lightness, red minus green, green minus blue). Particularly, we have success in practice with the HSV color system due to its close mimicry of how humans perceive and describe a color object, namely by its hue, saturation, and brightness. Hue represents the base attribute in the form of pure colors (red, yellow, purple, etc.); saturation provides a measurement of how much this pure color is being diluted by white light; and value/intensity is the amount of light (total darkness to total brightness) that a pixel appears to be reflecting. This representation is more intuitive than the RGB system and particularly useful when designing color-decoupling algorithms to resemble how humans observe, perceive, and describe the color aspect of objects.

Table 2.1 Haralick texture features

Entropy	$\sum_{i,j=0}^{N-1} P_{i,j}(-\ln P_{i,j})$
Contrast	$\sum_{i,j=0}^{N-1} P_{i,j}(i-j)^2$
Dissimilarity	$\sum_{i,j=0}^{N-1} P_{i,j}\lvert i-j\rvert$
Homogeneity	$\sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2}$
Uniformity	$\sum_{i,j=0}^{N-1} {P_{i,j}}^2$
Energy	$\sqrt{Uniformity} = \sqrt{\sum_{i,j=0}^{N-1} {P_{i,j}}^2}$

Compared to color and texture, which are used to describe the internal attributes, object **shape** is described using external characteristics (boundaries and their numeric measurements). This is a big family of descriptors that ranges from basic measurements (e.g. curvature, diameter), to statistical moments (e.g. Hu moments [15]), to advanced transformations (e.g. Fourier descriptors [16]).

Table 2.2 Common shape descriptors used for visual feature extraction

Roundness	$\frac{area}{\pi(d/2)^2}$
Circularity	$\frac{4*\pi*area}{p^2}$
Solidity	$\frac{area}{convex\ hull\ area}$

Convexity	$\frac{convex\ hull\ perimeter}{p}$
Compactness	$\frac{\sqrt{4 * area/\pi}}{d}$
Form Factor	$\frac{4 * \pi * area}{p}$
Aspect Ratio	$\frac{\max(boxHeight, boxWidth)}{\min(boxHeight, boxWidth)}$

The **textural** features within an image or a region are also a common feature group in image processing and retrieval. Its approaches are generally grouped into three categories: (1) statistical, (2) structural, and (3) spectral [3]. They each specialize in a subdomain of quantitative description and determination of the subtle yet complex patterns of pixel arrangement inside an image. Haralick proposed the most popular *statistical* texture features in the early 1970s [4] as a set of second order statistics of gray-level co-occurrence matrix (GLCM), which is the representation of frequencies of co-occurring pixel pairs with specific position relationship as well as their pixel values. The relative position relationships are determined both by direction and distance. In order to make these texture features rotation invariant, the average values from all directions are calculated. The *structural* approach represents complex patterns with a combination of texture primitives. The *spectral* approach, mostly notable for its Fourier spectrum representation [17] as well as Gabor [18] and Wavelet transforms [19], addresses texture feature extraction by treating images as a 2D signal and transforming it to another coordinate system where certain

information is revealed and manipulated. This is not only useful in texture feature extraction, but also contributes to image filtering.

2.2.3 Content-Based Image Retrieval

As opposed to concept-based image retrieval, which takes as input the text-based information about images, content-based image retrieval analyzes the image content itself and/or the visual features derived from the content, rather than keywords, captions, tags, etc. It is especially useful when the textual information is not available or insufficient [20]. While the earliest application of content-based image retrieval (CBIR) can be dated back in the early 1980s [21], the fields entered its active phase in the 1990s [22] and slowly but steadily advances for the past 20 years.

The basic components of a typical CBIR system include: feature extraction, image storage and retrieval, similarity measurements, and graphical user interface (GUI). These components are designed and developed to interact with each other in order to achieve the successful content-based image retrieval.

2.2.4 Machine Learning Techniques

Inevitably, machine learning techniques would be utilized in the process of imaging informatics. Machine learning, a subdomain of artificial intelligence (AI), takes as input empirical data and learns the underlying patterns to build a modeled system that does not strictly process data with explicit rules but rather with induced knowledge to make predictions on provided data [23]. The model is "learned" using training data, which are explicitly labeled/annotated with predefined finite collection of categories (labels). The patterns of the underlying

mechanisms are induced based on labeled data. Such data-driven models will then be evaluated using test data sets that are also labeled but are treated as if their labels are unknown to the model. The accuracy of model prediction indicates the usefulness of learned model. Machine learning algorithms can be categorized based on different criteria. They are commonly grouped into two categories, *supervised learning* and *unsupervised learning*. The most typical supervised learning is a group of algorithms called classification. The unsupervised learning mostly involves the discipline of data clustering.

2.3 Summary

This chapter introduced a selective set of topics that are frequently referred to and seen applied in the domain of biomedical imaging informatics, or generally in imaging informatics. We will revisit their details in subsequent chapters when we discuss specific applications.

CHAPTER THREE

WEB-BASED BIOLOGICAL AND CLINICAL IMAGE MANAGEMENT

Starting from this chapter, we will present a series of case studies in multiple research domains under the umbrella of biomedical imaging informatics. First, we showcase several web-based image management systems that were designed and developed for biological and medical domains. For each application, we also demonstrate our approaches to handle both generic and domain-specific challenges in the perspective of managing biomedical images.

3.1 Problems and Challenges

With the advances in high-throughput imaging, biologists are gathering an ever-growing amount of high quality, high resolution, and high volume imagery data. However, the analytic tools to handle such large-scale image data are still trying to catch up with the rapid pace of data generation. To ensure the quality of image analysis and retrieval, a set of ground-truth data need to be constructed by manual annotations from domain experts. This is particularly painstaking for biological image data due to the nature of imagery media and subjects of interest. Thus, an efficient yet reliable annotation strategy needs to be developed to accommodate the fast pace of image analysis. Furthermore, lacking a systematic and collaborating system to study and test biological hypotheses that rely on the biological morphologies is in need of improvement.

Dermatology is one of the most visually oriented fields in medicine. Dermatologists rely heavily on clinical images to diagnose and treat patients, conduct research, and instruct residents and fellows. Moreover, clinical images

are the primary media for dermatology professionals to exchange knowledge and experience with peers and to teach residents and students. Dermatologists learned to individually manage their image collections over the years to accommodate their clinic works, research activities, and education duties. Yet, as the number of years of experience grew, so did the volume, complexity, and variety of medical images. Managing the ever-growing image collection became complicated and time-consuming for an individual to handle, let alone making it accessible within an entire practice.

Accessing to medical specialists is not always a convenient choice for patients and their families who live in rural areas and under-developed countries where medical conditions are limited. The store-and-forward telemedicine (SAFT) systems provide a time-saving and money-saving way to connect patients to specialists for consultation. Dermatology is one of the medical domains that adopt tele-consultation in their clinical practices. Most existing store-and-forward teledermatology (SAFT) advisory systems use primitive Internet technology, such as e-mails or simply websites to handle case submission and communications among medical experts. However, it is not a secure choice when considering patient confidentiality and expert comfort assurance. Thus, it is imperative to design and develop secure and easy-to-use SAFT systems to improve the quality and efficiency of tele-consultation, in not only dermatology but also other medical professional subdomains that frequently need examinations of images and videos of patients.

In the eye of producing a new health IT application, a successful adoption is the overall goal. However, it is equally essential to study the process of

adoption and to understand users' interactions with the application, finding both common and different usage patterns. The studies of human-computer interactions in health care settings [24, 25, 26] greatly suggest the importance of understanding user behaviors and their feedback on the merits and limitations of the current design, resulting in improvements and suggestions to achieve a successful health IT implementation. Moreover, a useful, efficient, and convenient medical image management system by the side of medical professionals would impact not only the quality of clinical care, wound management and patient outcomes, but also the depth and breadth of medical research and education. The better users' usage behavior patterns are studied, the greater the understanding of the essential driving force of a successful health IT application, achieving an increased aforementioned positive impact and influence in the field.

To address the aforementioned challenges, we present our accomplishments in designing and developing image management systems for biologists and medical professionals. First we introduce BioShapes.org – a web platform for researchers from diverse domains to collaborate on a common interest – biological shapes. Particularly, a close collaboration among scientists in analyzing mitochondria dynamics is presented, emphasizing image management and annotation strategy for large-scale biological image data. Next, we present two web-based systems that are designed for dermatology professional for online consultation (a store-and-forward teledermatology system) and for clinical image management and annotation. To understand the workflow and usability of such systems, we also conducted quantitative analyses on domain expert

communication patterns and user behavior patterns while interacting with the system.

3.2 BioShapes.org and Biological Image Annotation

BioShapes.org was conceived and developed to be a web platform for a multi-institute collaboration of scientists, biologists, and mathematicians from diverse research domains. Although seemingly distinct if not remote from each other, these domain experts share a common interest in the biological shapes and their contribution in understanding the underlying mechanisms that drive their research forward. The subjects of interests include: tropical pollen and spores, bat ears and noses, embryonic hearts in chickens, and mitochondrial shapes and dynamics. The leading quest of this collaboration is to answer the following questions with computational approaches:

- Are current qualitative morphological categories (e.g., taxonomic, developmental) real and consistent?
- How does biological form relate to function?
- Can we predict physiological function, phylogenetic relationships, or ecological role through shape?

Based on these questions, four working groups are formed from an NSF funded project involving eight institutions: Mathematical methods and computational tools (MCT), Biological case studies (BCS), Visualization and data management (VDM), and Dissemination, education, and outreach (DEO). My involvement was primarily designated for VDM (developed and maintained BioShapes.org website where the goals and the results of the BioShapes group are

publicized, and sample datasets and metrics are available for download) and BCS (collaborated in mitochondrial dynamics study and Neotropical pollen and spores shape study). In this section, we will emphasize one of our close collaborations with fellow scientist on mitochondrial dynamics.

3.2.1 Background on Mitochondria Dynamics

Mitochondria are essential membrane-bound organelles found in most eukaryotic cells. A key mechanism of mitochondrial shape change is through their fusion and fission. Mitochondrial dynamics is essential to the spatial and temporal control of their functions in response to changing needs of dynamic cellular processes. In particular, it is critical to neurons because of their highly polarized structure. Defects of mitochondrial dynamics have been strongly implicated in many neurodegenerative diseases including Charcot-Marie-Tooth disease, Parkinson's disease, and Alzheimer's disease [2]. Overall, however, how mitochondrial dynamics and mitochondrial function are connected at the molecular mechanism level remains largely unknown. In a step towards answering this question, this collaboration has developed computational techniques for quantitative characterization of shape and motion dynamics of mitochondria. We used these techniques to analyze mitochondria dynamics in axons of normal as well as degenerative neurons from third instar larvae of *Drosophila* - an organism commonly used to genetically model human neurodegenerative diseases.

To capture the mitochondrial dynamics, axonal transport of fluorescently labeled mitochondria were imaged in axons within segmental nerves of dissected

Drosophila 3rd instar larvae with microscopic camera in the form of videos (time lapse sequences of images), lasting 60 seconds at a 3 second interval (Figure 3.1A). Next, individual frames are processed to identify mitochondria objects that are then tracked using a kymograph-based single particle tracking algorithm (Figure 3.1B) that was initially developed for tracking vesicles [27]. Each object is then labeled across all frames in a movie.

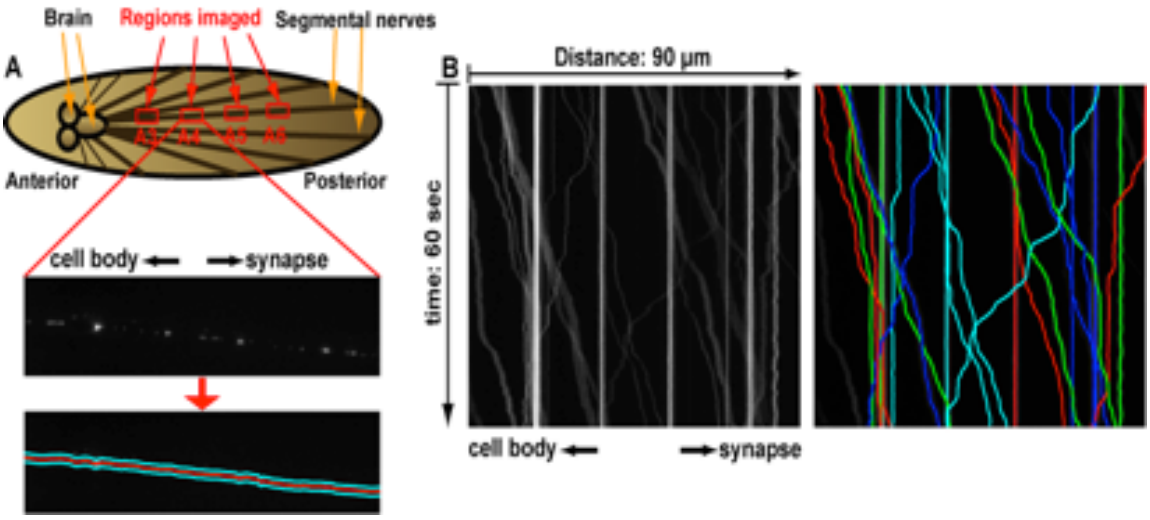


Figure 3.1 Imaging mitochondrial transport in segmental nerves of *Drosophila* 3rd instar larvae. (A) Four regions on a larva model (top) were imaged as videos in which each frame (middle) captures mitochondria objects within a highlighted (bottom) axon band where vesicle transport takes place. (B) Kymograph is generated by combining all frames vertically (left) and then recovered mitochondrial transport trajectories are artificially colored for later analyses.

A set of visual features are calculated to represent object morphology: area, orientation, extent, perimeter, convex hull area, solidity, eccentricity, major axis length, minor axis length, and equivalent diameter. Using these low-level visual features, different mitochondrial shapes are classified into 8 categories (examples are shown in Figure 3.2).

Bead: shapes that are approximately circular. Small mitochondria with size close to the diffraction limit are generally classified into this category.

Beadstring: shapes similar to a string of beads.

Rod: shapes with approximately uniform width and a high ratio of length versus width.

Pear: shapes with one round end and one tapering end.

Horseshoe: round shapes with a notch on one side.

Symmetric Oval: shapes that are close to ellipse with smooth ends.

Asymmetric Oval: shapes that are close to ellipse but with less even ends.

Irregular: shapes that do not belong to any groups above.

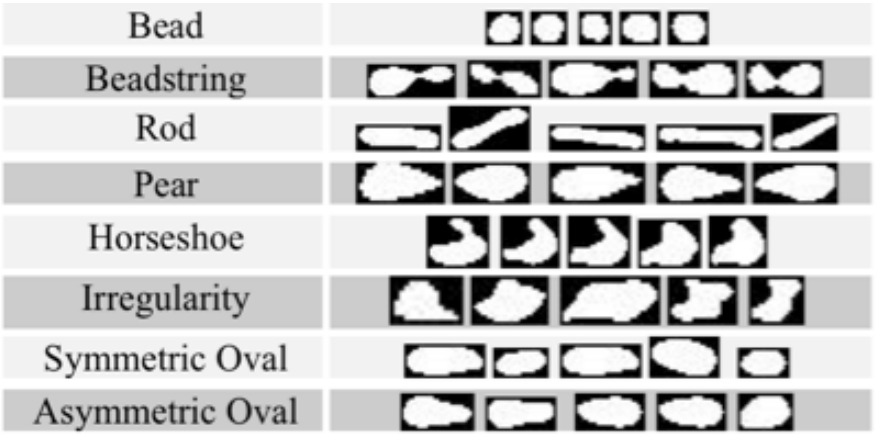


Figure 3.2 Examples of mitochondria shape categories

With each mitochondria (object) being detected and measured, trajectories of the mitochondria were reconstructed across the whole frames. Depending on the number of trajectories and number of actual frames that contains detected regions, the total number of regions in one movie range from 248 to 2425 in our dataset. Since it is virtually impossible to assign shape categories to individual regions, an efficient and user-friendly web interface is

designed to carry out the semi-automatic, bottom-up and rank-based region grouping process followed by manual labeling to produce training data for the subsequent data analysis steps.

3.2.2 Region Grouping

The detected regions on frames of each trajectory are extracted and saved as individual images. The regions in each trajectory are clustered and sampled through an automatic *clustering* step followed by three manual grouping *stages* to create a training set for all movies.

Clustering – Within each trajectory, regions are clustered into indefinite number of groups using DBScan with calculated visual features [28].

Stage I – As shown in Figure 3.3, clustered regions are displayed in groups with a few left ungrouped for user to assign to a group. In addition, user can create new groups if the clustering results are not satisfying.

Stage II – After all the regions are assigned to specific groups, users will be prompted to stage II where regions are sampled from each groups from all trajectories within a movie. As shown in Figure 3.4, the sampled regions will be grouped further into groups according to users' preference on morphological characteristics.

Stage III – After all movies are processed, the grouped stage II regions will be sampled on stage III and selected to be labeled in the labeling step.

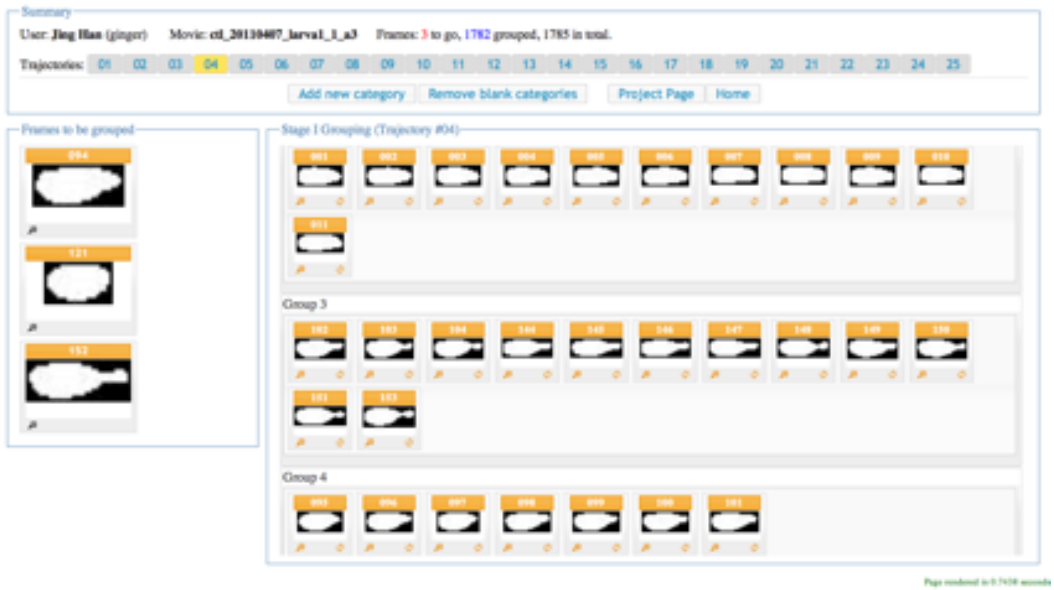


Figure 3.3 Step I of region grouping. In the shown page of a movie, there are 25 trajectories with 1785 regions extracted. Regions in trajectory #04 are clustered into four groups automatically based on their visual features.

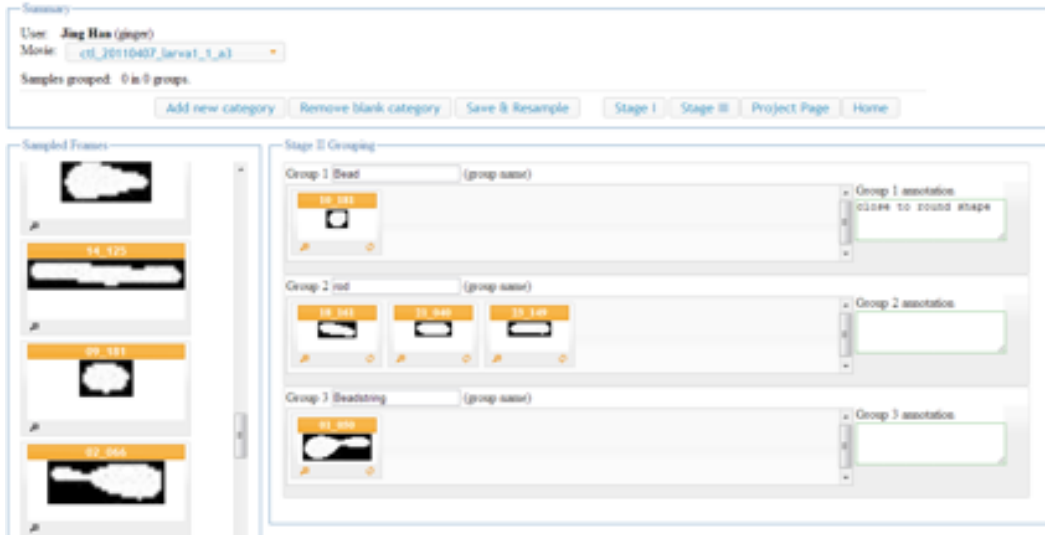


Figure 3.4 Step II of region grouping. Regions are randomly sampled from groups created on Stage I (one per group) and listed on the left panel for experts to drag-and-drop to the groups on the right panel. Sample list can be regenerated until experts are satisfied with the results.

3.2.3 Region Labeling

After previous three stages of grouping, at least one region will be selected from each group onto labeling step. Selected regions from stage III are listed on

the labeling website (Figure 3.5) for users to review and assign predefined semantic labels. In this study, we provide 8 shape labels and 3 size labels (small, medium and large). Consensus labels are used to label the rest of the training data.

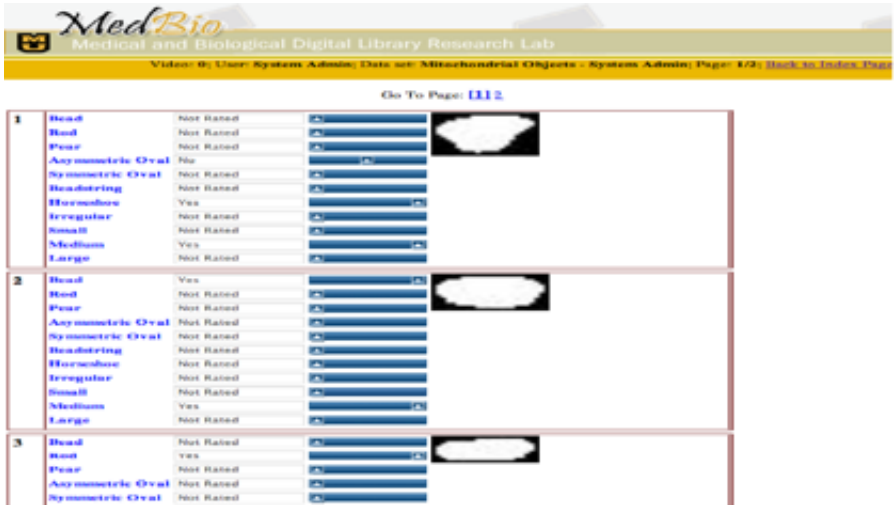


Figure 3.5 Region labeling webpage. Users use rating bars to assign labels to each region with three options (Yes, No, and Not Rated). Only one shape label and one size label are assigned to each region.

3.2.4 Semantic Labeling

For semantic modeling we generate associations between feature subspaces and domain semantic of interest. For example, in our experiments, we determined that 90.5% images that have the measurement in the feature subspace formed by

$$F6 \in [0.053, 0.749] \wedge F7 \in [13.09, 17.08]$$

were labeled “Bead”. Due to this high density we can create predictive association

$$\{F6 \in [0.053, 0.749] \wedge F7 \in [13.09, 17.08] \rightarrow "bead"\}$$

that can be used to predict the relevance of new, not evaluated images to the semantic “bead”. In our experiments we generated relevant subspaces using the

Apriori algorithm. To further refine the semantic assignment, the discovered associations semantic modeling (*SM*) is reduced to improve mean average precision (MAP) score of ranking. (Figure 3.6) After reduction, the *SM* model is used for prediction. For example, the semantic model for the semantic "beads" contains 18 association rules that segment the feature space using between one and three features. For details of this subsection, please refer to previous works in [29].

Figure 3.6 shows the results of ranking images by semantics using 9 semantics of interest (six from Figure 3.2 and three size-related categories). The dataset contained also four images labeled "horseshoe" that were used in data mining. However, a semantic model was not generated for this semantic due to lack of sufficient data. For this experiment, we have mined associations using the following *Apriori* parameters: minimum support 0.75% and minimum confidence 60%. As seen in Figure 3.6, these semantics return good MAP scores, demonstrating potential for predicting new semantic assignments.

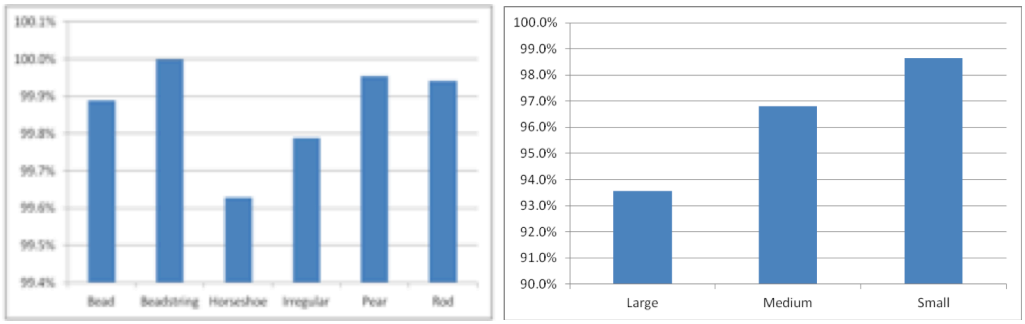


Figure 3.6 Results ranking images by semantics using *MAP* scores.

Data analytics for complex biological data require new ways of reasoning of the low-level features so that they can be associated with high-level biological meanings for better understanding of underlying mechanisms for scientific

discoveries. In this study, computational methods have been developed to mine the patterns of shapes of biological objects, automatically annotate the biological semantics of objects, present object dynamics in a computational way, and make the information searchable for in-depth studies. The success of the work will bring new informatics tools for the life sciences community to look into the dynamics of biological objects in a systematic and analyzable means.

Another close collaboration with palynologists on searching Neotropical fossil pollen and spores images based on shape characteristics will be highlighted in Chapter 5. Next, we will introduce two web-based applications in the medical domain, specifically, dermatology and teledermatology. However, we argue that our system and approach of handling clinical image and data management can be further adapted into other similar medical domains that, like dermatology, rely heavily on clinical images.

3.3 SAFT for Dermatologists and Patients

3.3.1 Background on SAFT and Ichthyosis

In the domain of dermatology consultation, the most traditional fashion is, as the majority of other medical specialties, relying on face-to-face encountering where dermatologists would both visually and physically examine patient's skin lesions. However, due to the limitation of money, traveling, accessibility, and resource allocations, not all patients have the convenience to receive a dermatological consultation when they most need it in a timely fashion and their primary care doctors do not have the specialty to diagnose and treat the disease. This is when telemedicine, specifically teledermatology, fills the gap in medical

services. Telemedicine, as defined by American Telemedicine Association [30], "is the use of medical information exchanged from one site to another via electronic communications to improve patients' health status". It can be utilized as a bridge between primary care doctors and experts with specific medical knowledge and experience, allowing evaluation and treatment of difficult medical cases through telecommunication technology. In dermatology, telemedicine has already played a crucial component in delivering efficient service of diagnosis and management of dermatologic diseases for patients and also providing advisories for physicians in primary care settings. In the United States, teledermatology has been used to improve access to care in rural and medically underserved areas [31].

There are two general types of telemedicine in the field of dermatology: (1) real-time tele-consultation between patients, accompanied by their primary care providers, and a dermatologist located afar; and (2) store-and-forward teledermatology (SAFT) system where medical cases are submitted by primary care providers into a central system and later attended by distant experts when they are most available. During the telemedicine process, communication happens over the web or via email asynchronously. Moreover, it makes it possible to take advantage of remote experts who will have time to carefully study difficult cases before making a medical decision. Most existing SAFT advisory systems use primitive Internet technology, such as e-mails or simply websites to handle case submission and communications among medical experts. However, it is not a secure choice when considering patient confidentiality and expert comfort assurance.

Ichthyosis refers to a group of inherited skin diseases characterized by dry, thickened, scaling of the skin. Affected patients often report difficulty in finding physicians who are knowledgeable about their conditions and their treatment. F.I.R.S.T. [32] is a support organization for Ichthyosis patients and their families that have pursued the use of store-and-forward teledermatology to facilitate communication between physicians caring for patients with Ichthyosis and experts in these rare diseases. Since 2009, we have been providing a web-based tele-consultation platform for primary doctors, Ichthyosis experts, and patients as well as their families this community. With steady growth and improvement over the years (Figure 3.7), it has become well accepted by not only primary doctors but also patients who suffer from this rare skin disease from around the world. There have been over 120 medical cases submitted by primary care doctors and social workers, that were then reviewed by over 30 specialists. The general purpose for this system is to provide a secure and easy consultation environment for experts to discuss dermatological cases submitted worldwide.

The research team here at the University of Missouri conducted an analysis on Ichthyosis experts' activities across two years of consulting teledermatology cases on this platform to discover behavioral patterns, which could be used to provide feedback to users for future involvement and improve the development of new features and workflows for existing system and other similar tele-consultation systems.

Figure 3.7 Two generations of F.I.R.S.T. Tele-Ichthyosis websites.

3.3.2 System Structure

In our system, the teledermatology process has been broken down into several distinct modules. From the system architecture shown in Figure 3.8, the primary modules for case consultation include submission, discussion, final report, voting, and feedback.

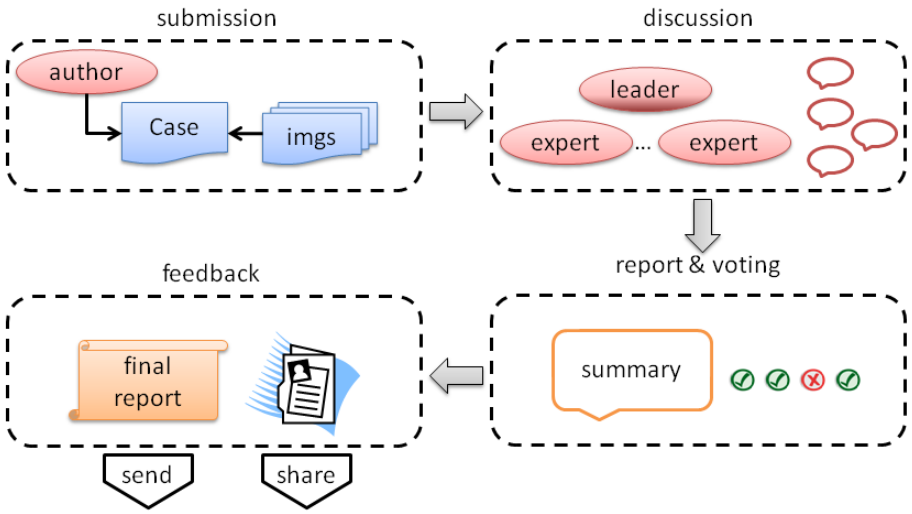


Figure 3.8 Tele-Ichthyosis system structure.

Case Submission: During submission, the case author is required to provide crucial information related to the patient's condition, including brief description of medical history, past treatment, symptoms and medication, etc. Clinical images may be uploaded allowing participating experts to view pertinent

findings. After creating the medical case, the author submits the case for approval by one of the system administrators, who are the moderators responsible for the daily functioning of the SAFT system. If the administrator considers the case description insufficient, the author will be asked to provide additional information or make modifications.

Case Discussion: Once the case is suitable for discussion, it is assigned to an expert and he/she will select several other experts for case discussion. Each expert submits comments to the forum until a satisfactory decision can be reached. If it is determined that more information is required, the leader will correspond with the author to provide additional details. Because all correspondence with the author goes through the case leader, we protect the identities of contributing experts so that they will feel more comfortable providing feedback in the forum setting.

Case Report: Once the case leader determines that the discussion is complete, he/she will compose a final report to summarize the main points including disease concept explanation, diagnosis (if reached), and treatment suggestions to be sent to the case author.

Case Voting: Before sending out the case report, all of the participating experts will be asked to approve the summary through a voting system. The confidential voting results will be sent out along with the final report to the author.

Case Feedback: As the last stage of the teledermatology process, the case leader will close the case, compile the final report and voting results, and provide the feedback to the case author. The final report will be sent to the case author

without leaking the identities of participating experts and the closed case can be shared within the system to fulfill the education purpose.

3.3.3 Case Variables

Each case consultation involves a case author, a case leader, and several participating experts. The number of participants, comments, additional pictures and case request types vary among cases existing in the system collection. The request type for a submitted case can be asking for differential diagnosis, treatment and management, or a general purpose of discussion for an interesting case. We use those basic variables as observational evidence to extract common patterns across the whole collection of cases.

3.3.4 Comment Linearity

It is not uncommon to see that in any forum-style discussion some comments are a direct response to a previous comment. The interactions between participants in commenting are grouped into two levels – comments with direct-responded targets are level 2 while those that simply introduce new inputs are level 1. A measurement of these leveled comments is defined as comment linearity L - the ratio of level 1 comments and total comments, $L \in (0, 1]$.

$$L = \frac{\#level1\,(linear)\,comments}{\#all\ comments} \tag{3.1}$$

A case with only level 1 comments ($L = 1$) is considered a linear discussion case while the one with level 2 comments will have smaller L value indicating a less linear discussion.

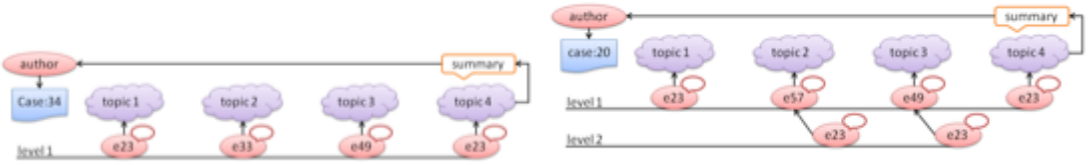


Figure 3.9 Comment linearity examples. Left: a case with L= 1 and no image; Right: a case with L = 0.67 and 10 images.

3.3.5 Communication Modeling

The interaction between experts and cases is modeled using the concept of social networking, which can systematically identify central users as well as those who remain isolated by constructing a graph (Figure 3.10). In such a network graph, composed of nodes and edges, a node represent a participating expert and an edge represents an instance of communication between experts. Consequently, we can observe the density of a node by counting the number of connections. This gives us a sense of the level of interaction for participating experts.

Expert nodes have various numbers of connections (in red), reflecting the number of cases to which each expert contributed. Expert 23, the central node in Figure 3.10 has the largest number of collaborations, and acts as the hub of the social network. Other experts are associated with expert 23 through case collaborations. A hub node has an important role in a social network, as it frequently serves as an intermediary between unconnected nodes. As we examine expert nodes, we noticed nodes 43 and 4 that have only a single link. Similarly, case nodes 45 and 47 also have one edge indicating a singular collaboration with an expert. Such aspects provide a valuable insight into potential reasons why either experts or cases are isolated. The case leader plays a critical role in a

fruitful and ultimately successful discussion. A successful leader requires good understanding of his/her obligation, familiarity of key functions of the system, and proactive leadership qualities for moderating the consultation process. The final feedback is composed by the case leader collecting key points from all the comments and is sent to the case author. Case authors benefit most from responses that contain insights based on the collective experience of the experts and rich medical analysis and explanation. Sometimes the inquiry problems are not fully addressed in the final responses because of lack of consensus form all participants. Those responses are considered as weak responses. There are 42% of cases concluded with weak final responses. However, the most active leader participated in seven cases with only two weak responses.

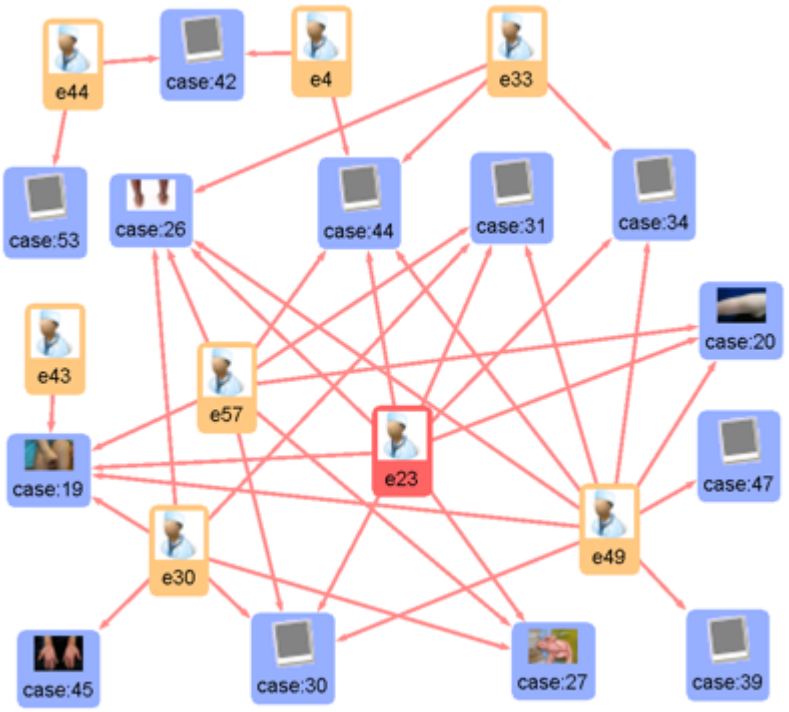


Figure 3.10 Communication networking around a core expert (hub).

Based on our analysis of cases in our teledermatology system, we have arrived at the following suggestions for the design of telemedicine systems and potential improvements of our existing system:

1) Certain types of cases would benefit from a customized workflow, such as simplified process for cases that can be answered by a single expert.

2) Users can be identified and targeted for notifications or particular cases through social networking tools.

3) Users can be asked to categorize their own cases for more specialized workflows for more suitable discussion group selection to lead to a better and faster discussion.

4) Users should have a mechanism to provide a follow-up and feedback about the usefulness of expert's suggestions and system usability.

5) Experts should be strongly encouraged to communicate within the system.

6) Social network graphs can identify critical experts and ensure robust and healthy communication in the event of an experts' absence.

3.4 Clinical Image Management and Usability Study

3.4.1 Background

The current electronic solutions to manage clinical images can be loosely categorized into three types: customized modules for commercialized EMR/EHR systems, stand-alone software applications for desktop computers, and web-based applications or resources. Each type has their own strength to support health care professionals. While each has characteristic merits on providing

services to medical professionals, there are also some inherent aspects. For example, modules from EMR vendors or a separate product can be relatively expensive and rigid in customization; stand-alone applications may require certain operation system configurations; and free third-party software are conceived and developed with the purpose of serving general image management (for example, Picasa, iPhoto) in the core design concept. However, web-based applications or resources are designed for multiple professional users from the entire department or private practice in a secured environment. With a secure Internet connection, there is no limitation of user location or choice of web browsers. Furthermore, there is no software installation needed on any desktop computers. Table 3.1 lists some well-known web-based resources for dermatology, including our own MDID system.

Table 3.1 Web-based dermatology image resources.

	EMR Assoc.	Image Browse	Simple Search	Multiple Search	Additional Info.	Personal Collection
DermNet.com	✗	✓	✓	✗	✗	✗
DermNetNZ.org	✗	✓	✓	✗	✓	✗
DermQuest.com	✗	✓	✓	✗	✓	✓
DermAtlas.org	✗	✓	✓	✓	✓	✗
DermAtlas.net	✗	✗	✗	✓	✓	✗
DermIS.net	✗	✓	✓	✗	✓	✗
Dermo-Image	✓	✓	✓	✓	✓	✗
DermaShare	✓	✓	✓	✓	✓	✓
MDID	✓	✓	✓	✓	✓	✓

Mizzou Dermatology Image Database (MDID) is designed as a web-based, database-driven clinical image management system for dermatology professionals at Department of Dermatology, University of Missouri and its

affiliated clinics. This system has also been perceived as a generic model for any dermatology practice and some other image-intensive subspecialties. MDID provides daily image-involved routines, such as upload, view, organization, sharing. All of which take place in doctor's offices, patient rooms, and nurse's workstations. The main purpose of MDID system is to serve as an easy-to-use, secure and efficient interacting media between human (dermatology professionals) and machine (web and database servers). Instead of directly managing image files on the file server, users would access the designed MDID interface components to achieve clinic image management tasks (Figure 3.11).

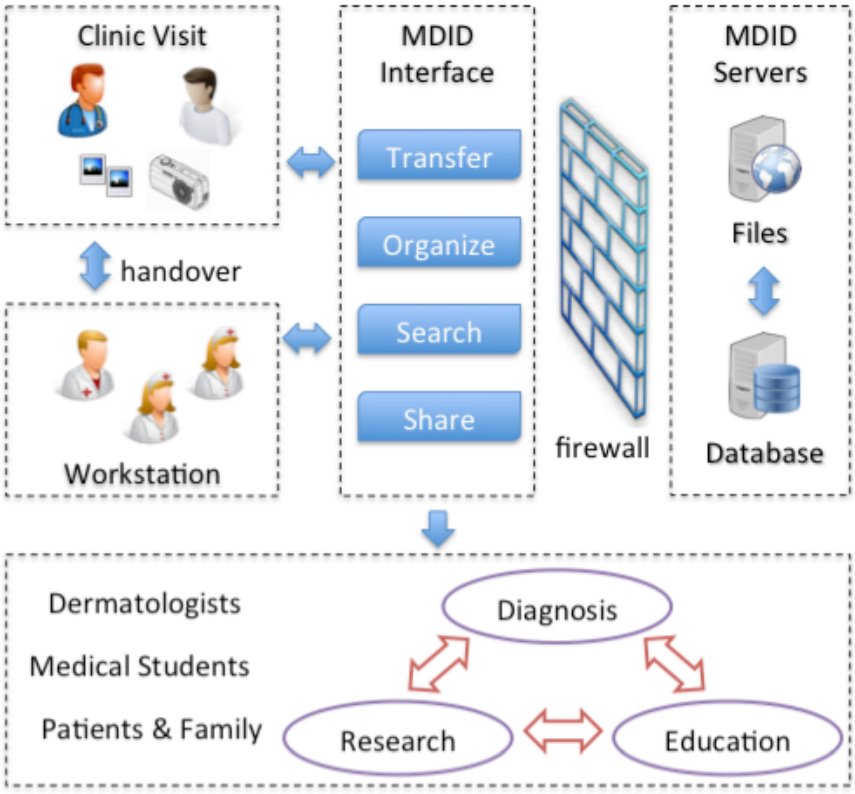


Figure 3.11 System structure of MDID.

With MDID, digital images are transferred to and stored on a web server, while all relationships among patients, images, and dermatologists, etc. are

stored and managed in a separate database server. Instead of letting users directly manipulate images on the file server, MDID provides all the necessary functionalities through its web interface, where images are handled with designed standard procedures. The merits of separating users from direct data access are to protect data integrity, to prevent image duplication and accidental deletion, and to monitor image access.

The core entities are associated around clinic visits. In other words, one specific clinic visit, taking place at a clinic, involves a patient, one or more dermatologists (physician), and possibly some images. Each image can be annotated by a DermLex™ concept as its diagnosis and multiple free-text tags. Another convenient function that MDID provides is folders. Images can be put into multiple folders stored only in the database according to user's preference just as what they may prefer in their conventional organization of images. In this case, managing images is still customized but images are not duplicated in multiples physical file instances all across the disk space. MDID records all web usage activities in its log system for application usage analysis and data recovery. The rich log data are intensively analyzed in our study to discover and understand how dermatology professionals utilize MDID into their clinic routines on a day-to-day basis.

Beyond the implementation of a useful web-based application for clinic image management, another goal of this work was to study the process of adopting a new health IT application in a health care specialty and to discover interesting user behaviors throughout the adoption for their professional training, diagnostic activities, and academic research. In the next section, we will

emphasize our longitudinal study of system usability and usage behavior patterns with MDID as a general model for other similar systems.

3.4.2 Data Collection

All professionals in the Department of Dermatology and its clinics, who work closely with clinic image management, are the targeted subjects. For example, nurses and technicians would upload images to MDID; dermatologists would search for patients and images; and any approved employee associated with the department is able to retrieve images. Our longitudinal study is designed to utilize several research instruments.

- **Online surveys** were distributed three times: a pre-launch survey to collect participants' perception and attitude toward their image-related routines before MDID implementation and two periodic user feedback surveys amid and at the end of the study period.
- **Interviews** were conducted twice at around the same time with the two periodic surveys to catch participant's verbal description of their experience with MDID.
- ***Field Observations***: The department holds weekly research sessions to exchange clinical experience, case progress, and discoveries. The research team attended these sessions as sit-in audience members and took notes of any activities that involved medical image management.
- ***User Activity Logging***: To avoid disrupting the user's experience, a built-in logging module of MDID was used to capture the application access activities. Unlike the typical web log analysis that uses server

logs, MDID's log module records not only standard web page access information, but also user-specific information related to MDID's functionalities.

3.4.3 Usage Log Data Analysis with Sequence Mining

Log data are recorded as time sequences of events triggered by user actions and web application functions. Therefore the analysis of log data is a study of discovering salient sequential patterns. The raw log data records generated by the MDID log system store necessary information in regards to individual actions triggered by mouse clicks. A *session* on MDID consists of a series of mouse-clicks, with each mouse click producing a list of ordered elemental actions with the data structure shown in Table 3.2. A single action example shown in Table 3.2 tells us that on date '2013-07-26' at time '16:39:20', after viewing a record page of patient with medical record number (MRN) '99-99-99-99-9', user 'jsmith' chose to see all visits pertaining to this patient. This single action is coded with *i10* and is uniquely identified with Log ID 371611. Additional information of this page access shows that this patient has an ID of 4 and attended 2 clinic visits where 12 images were captured.

Every page loading is triggered by a mouse-click that subsequently initiates a series of elemental actions to complete a specific task. Different combination of such elemental actions can form different tasks. We were interested in how the tasks were performed with what kinds of patterns. Therefore, two rounds of necessary mappings of the raw log records were conducted. The first round of mapping assigns a unique code to each type of

elemental actions, followed by the second round of mapping assigning unique tasks that are composed of such actions. The coded series of tasks after the second mapping are eventually used for sequential pattern mining introduced in the following section.

Table 3.2 Essential attributes of log data structure in this study.

Attribute	Description	Example
Log ID	Unique identifier of the log instance	371611
Timestamp	Time at which the instance happened	2013-07-26 16:39:20
User ID	User who made the action	jsmith
Action ID	Functions or Page Access	*i10*
URI	Current accessed page	Images/patient/99-99-99-99-9
Referrer	Page that precedes current URI	patients/show/99-99-99-99-9
Note	Additional action-specific information	[patient_id]=>'4';[date count]=>'2'; [image count]=>'12';

3.4.3.1 Sequential Pattern Mining

Sequential pattern mining, or sequence mining, is a data mining technique that discovers sequentially interesting information with statistical significance. It is a popular choice of many applications to gain knowledge of research fields, such as consumer behavior, biological sequence analyses, and web usage analysis. Agrawal and Skikant [33] first defined the method in year 1993 as one to find all frequent subsequences, in a set of sequences, with occurrence frequency no less than user provided minimum support threshold. In most scenarios, the problems are described symbolically; that is, a set of items are defined to form sets of events that are then sequentially ordered into sequences. A sequence database consists of a set of sequences uniquely identified by their IDs. The sequences whose support values are no less than the minimum threshold are considered to frequently represent interesting sequential patterns in the database. The MDID

log system records every action initiated by users' mouse clicks. To understand how its users interacted with different functionalities and how often various usage patterns occur at different periods of time, a sequential pattern-mining algorithm, SPADE (Sequential PAttern Discovery using Equivalence classes) [34], was utilized on collected historical user log data over the course of this study.

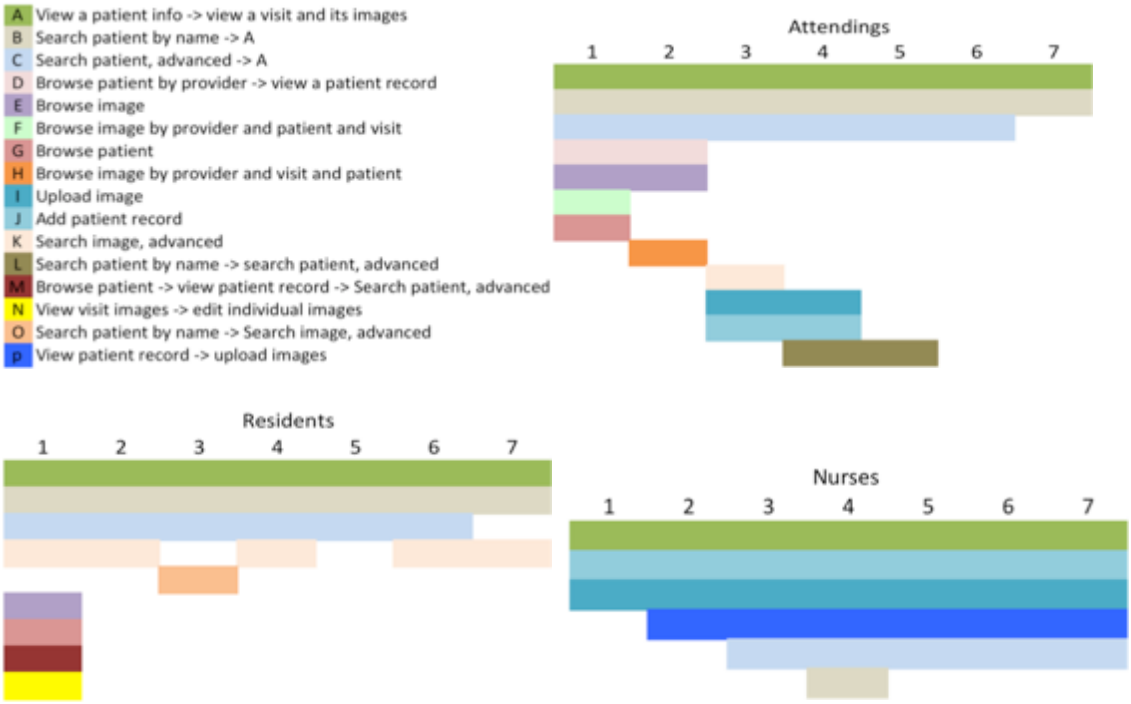


Figure 3.12 Frequent usage patterns discovered over time windows for each user group.

3.4.3.2 Recommendations

Once users' usage patterns are discovered, a better understanding of users' needs are revealed. Consequently, a planned strategy can be suggested to further improve user experience, system efficiency and effectiveness.

- The most frequently used functions should be made most easily accessible and tailored to different groups of users. For example, a

user's role is recognized as soon as they log in and a customized dashboard is constructed to display those most accessed functions.

- There could be two strategies to deal with those functions that are less likely to be used by certain groups of users.
  - Place them in a less distractive location on the application interface, so that users can focus on what are most useful to them.
  - Meanwhile, periodically prompt users with those less often used functions and encourage them to try out.
- Introducing a new health IT application to a well-established workflow needs not only one but several tutorials and assistance along the adoption process. A close collaboration between developers and end users will encourage better and faster adoption.

Although at this point, MDID is designed only for dermatologists from one department, its design concept, data structure and functionalities all make it possible to be adopted into other clinical domains. Furthermore, patients are not left out of the picture. Doctors have shared their experiences that at the clinic, patients welcomed the in-time observation of their historical images and they gained better understanding of their disease situation and the importance of careful management. It is obvious that users usually have a targeted patient in mind before they initiate the searches for either images or patients. This discovery coincides with MDID's designed purpose.

3.5 Summary

In this chapter, three web-based image management systems were developed for biologists and dermatologists for a broad usage: image storage, annotation, tele-consultation, clinical case studies, educations, and more. We demonstrated our approaches to handle large-scale biological image organization and annotation with efficiency (hierarchical grouping) and meaningfulness (semantics discovery). This can be extended many other applications in the fields of biological image analysis where large collections of similar images can be grouped and then annotated both automatically and manually.

We also demonstrate that a useful web-based resource system would assist medical professionals to perform tele-consultation around the global providing most-needed medical advisories and to manage day-to-day clinical images with security and efficiency.

Another related research work was also conducted on a mobile extension of MDID system. We studied the adoption process of mobile MDID on iPads that functions as a complimentary method for tele-conference consultation sessions between dermatologists at University of Missouri Health Center and local doctors and their patients in rural mid-Missouri.

CHAPTER FOUR

VISUAL CONTENT EXTRACTION

In this chapter, we shift our focus from the front-end of image management to the backstage where advanced image processing and analysis approaches play their roles in extracting visually meaningful content from biological and medical images.

4.1 Problems and Challenges

Scientists and researchers who work closely with digital images obtain high-level and domain-specific knowledge through extensive training and years of experience in the related fields. They make discoveries by visually examining the content of images. As the volume and speed of digital image generation increase, so does the need to develop accommodating computer programs to process and analyze images with both quality and efficiency. One of the toughest jobs is to program computers to examine images as human. In order to do so, we first have to understand the reasoning that leads to knowledge discoveries. Next, we will need to expressively design computer programs to automate such reasoning process. Therefore, it is imperative to develop smart visual content extraction approaches.

In digital image processing and analysis, a series of processes need to be designed and developed to work together and eventually extract the visual content residing inside images. They usually involve image pre-processing, object segmentation and identification, and visual feature extraction. Domain-specific knowledge drives the design logic of actual computer vision programs for

automatic image processing and analysis. Therefore, a comprehensive understanding of research domains is crucial to provide helpful analysis tools for researchers. In the following sections, we will introduce several biomedical imaging fields and their specific needs for automatic image processing and analysis followed by our approaches to address such needs.

4.2 Grain Shapes of Neotropical Pollen and Spores

4.2.1 Background on Palynology and Grain Shapes

Palynologists use the morphological characteristics of pollen and spore grains to identify, classify, count, compare and log plant diversity within geologic samples from different geographical locations and ages. These data are used to address research questions in areas such as biostratigraphy, paleoecology, biodiversity, climate change, taxonomy, evolution, and are even increasingly employed in forensics. The potential sample size represented by a fossil pollen sample can be very large, since hundreds to thousands of grains can be preserved in a drop of pollen residue extracted from a geological sample (rock or sediment); but, the classification of samples is still primarily qualitative and manual, based on the visual identification of key morphological features, and requires significant experience and expertise [35, 36].

This manual, intuitive approach to classification [37] potentially results in discrepancies in taxonomic identifications due to individual differences in analysts' interpretation of morphological details, familiarity or experience with a given suite of taxa, fatigue, and preservation of fossil pollen material. Morphological similarity among related taxa may also decrease the taxonomic

precision of identifications, due to the inability to observe or to define morphological differences [38]. Moreover, the intrinsic morphological variability found within pollen grains from even the same species makes it difficult to assess the morphological boundaries of any given fossil species. There are few published studies of how much morphological difference can be consistently recognized among analysts [39]. As a result, the recognition and formal naming of new morphotypes relies on a certain degree of consensus from a community of experts. However, with advanced imaging technology, digital microscopic pollen images are being generated with increasing speed and volume, producing opportunities to improve upon the traditional manual identification and sorting of grains and to produce higher throughput approaches to pollen analysis.

There are prominent databases and software applications in literature developed to assist palynologists in their identifications. For example, Bush and Weng [40] designed a downloadable Neotropical pollen database as a freeware for Neotropical palynology researchers. It provides multiple-access keys to query the database with flexibility and tolerance in missing data attributes. The collection contains pollen images, primarily taken with transmitted light microscopes, from more than 1000 Neotropical species. Morphological features, such as pollen shape, pore shape, reticulum shape, and pollen size, can be used to query the database. A second pollen image database, PalDat [41], has a similar query structure and web-based interface and includes both transmitted light images and scanning electron micrographs (SEM) from ~2200 modern species and 32 fossil ones. Neotoma Paleoecology Database [42] is another example that provides complex information for Pliocene through Holocene mammals and

fossil-pollen data from published literature and collaborating individuals from multiple institutions. Its main purpose is to map spatiotemporal taxa distribution [43]. Classfynder developed at Massey University is a stand-alone system that provides a framework for image acquisition and classification of modern pollen materials [44]. Its experiments suggested that computer performance in pollen identification and classification was comparable to human experts, but with better consistency. This work can be further extended to image search using extracted visual features of identified grains in both modern and extinct species.

While these image databases and software applications serve as valuable resources for pollen identification, having to manually label and compare morphology is both time-consuming and subject to the idiosyncrasies of individual analysts. Automated visual content extraction allows analyses to be kept more consistent across multiple sites, and is especially useful when there is an unknown sample with new morphotypes that needs to be compared against existing collections. Previous applications of machine-based classifications for pollen identification have focused on the accuracy of the end classification [45] and generally do not provide a mechanism for establishing the community-level consensus of identifications that is required when working with extinct species. Developing a broader platform for capturing and sharing expert knowledge builds on previous machine learning and image database efforts and provides a pathway for making these tools widely accessible.

The ultimate goal of this research is to use informatics tools to assist palynology study in form of increases in speed, efficiency and reduces in inter- and intra-observer inconsistency and labor intensity and eventually to determine

species of new samples. In the steps toward this goal, we have developed a database-driven application that integrates the analysis of image content, grain object morphology, morphology semantic modeling and annotation, and user-computer interaction through web pages for multi-modal information integration. To our knowledge, our work is the first attempt to develop a unique search engine to utilize image-based morphological content for grain image retrievals in palynology. In this chapter, we will be emphasizing on the image processing and analysis components of the project while leaving its content-based image retrieval aspect presented in the next chapter.

4.2.2 Data Collection

In this study, 525 images from Miocene-aged pollen and spore material were taken from a stratigraphic section of Falcon basin in Venezuela [46, 47]. These images represent the 15 pollen taxa and 5 spore taxa listed in Table 4.1. Morphological information for each of the taxa was collected from the Smithsonian Tropical Research Institute (STRI) palynological database [48], which contains the morphological descriptions of ~2700 species of Neotropical fossil pollen and spores. Images were taken using a Zeiss AxioImager microscope, Plan-Apochromat SF25 (63×, 1.4NA, oil immersion) lens and a Zeiss AxioCam ICc 3 digital microscope camera. This subset of taxa was selected based on its morphological diversity and sample availability at the time of study. Since overlapping of grains and debris is not uncommon in prepared microscopic slides, each sample image was cropped roughly at the center of a grain without intentionally avoiding debris.

Table 4.1 Dataset details of Neotropical pollen and spore samples.

	ID	Taxon	# Grains	# Images (mean)
Pollens	1014	*Clavainaperturites microclavatus*	6	24 (4.0)
	148	*Clavainaperturites clavatus*	7	22 (3.1)
	246	*Echiperiporites estelae*	5	18 (3.6)
	1430	*Echiperiporites scabrannulatus*	7	24 (3.4)
	365	*Grimsdalea magnaclavata*	5	21 (4.2)
	254	*Malvacipolloides maristellae*	7	25 (3.6)
	450	*Mauritiidites franciscoi var. franciscoi*	9	46 (5.1)
	451	*Mauritiidites franciscoi var. minutus*	7	34 (4.9)
	511	*Perisyncolporites pokornyi*	7	18 (2.6)
	552	*Proxapertites psilatus*	7	29 (4.1)
	570	*Psilamonocolpites medius*	7	35 (5.0)
	571	*Psilaperiporites minimus*	5	19 (3.8)
	688	*Retitrescolpites? irregularis*	9	32 (3.6)
	722	*Retitricolpites simplex*	7	24 (3.4)
	767	*Rhoipites guianensis*	7	26 (3.7)
Spores	43	*Echinatisporis muelleri*	7	28 (4.0)
	45	*Magnastriatites grandiosus*	7	24 (3.4)
	282	*Kuylisporites waterbolkii*	7	25 (3.6)
	44	*Crassoretitriletes vanraadshooveni*	6	28 (4.7)
	46	*Polypodiisporites usmensis*	5	23 (4.6)

4.2.3 Grain Segmentation

There are multiple options of image analysis toolkits [49, 50, 51] to roughly segment a centrally placed object from a field of view [3, 52, 53, 54]. We used several of these methods to extract grains from the image background. More refined methods were then conducted to extract morphological features. We detail the process below.

The cropped images of individual grains are RGB color images. However, in computational analysis and machine vision research, this color system is not

always the best configuration to represent how human observers perceive content and pattern. Therefore, we converted and separated the original RGB images into three single channel images using the HSV (*hue*, *saturation*, and *value*) color system [3]. In each image channel, pixel values not only represent part of the color space, but also contribute to segmentation of objects [55] and calculations, representations of advanced visual constructs, such as textural content and shape characteristics. Using only gray scale images limits the ability to segment objects of interest efficiently or extract underlying visual patterns that comprise the image content. While *value* images provided the viewer with detailed texture of the grain, *hue* and *saturation* images allowed us to discriminate between foreground objects and background. Since our goal in grain segmentation was to find reasonable contrast in order to recognize the grain contour, we merged the *hue* and *saturation* images to reconstruct an intermediate image that displayed better separation of grains from background using the following equation.

$$p_M = \left(\frac{p_H}{180} * w_H + \left(1 - \frac{p_S}{255}\right) * w_S\right) * 255, \qquad 0 \le p_M \le 255 \tag{4.1}$$

The merged image pixel value, p_M, is a weighted combination of pixel values from hue pH and saturation p_S images at the same pixel location. For example, Figure 4.1A is an image of a pollen grain (*Clavainaperturites microclavatus*) that is converted and separated into three single channel images (Figure 4.1B-D) using the HSV color system. The hue image (Figure 4.1B) and saturation image (Figure 4.1C) are then merged with weights w_H and w_S to produce an intermediate image (Figure 4.1E). We tested a sizable sample of

images using various weight combinations and observed an influence of weight choices on segmentation performance (Figure 4.2).

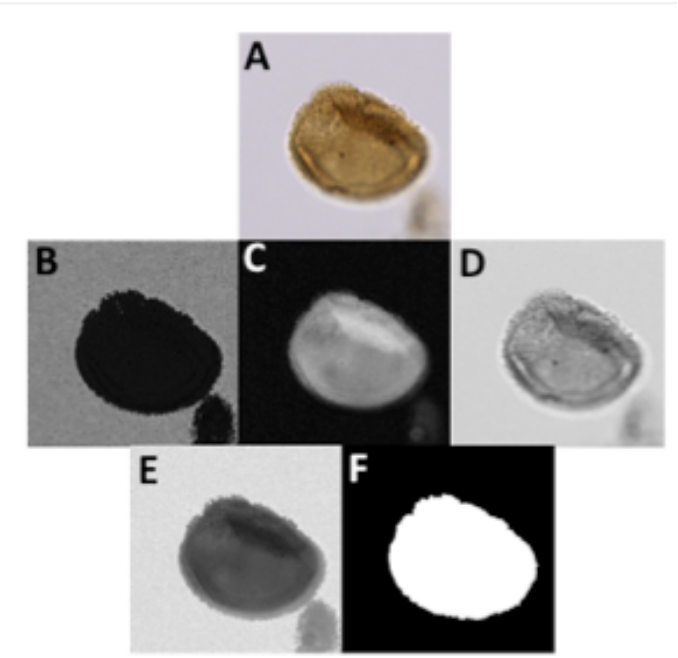


Figure 4.1 An example pollen grain (*Clavainaperturites microclavatus*) image segmentation process. The original RGB image (A) is converted from a single RGB image to three single-channel images—hue (B), saturation (C), and value (D). (B) and (C) are then merged using selected weights on pixel values (Eq. 4.1) to generate an intermediate image (E) for thresholding, morphology operation, watersheding, and connected component operations. This ultimately segments the main grain object (F) from the rest of the image, including background pixels, trivial particles, and debris.

The bigger w_H, the more the hue value was emphasized; therefore, image pixels were separated based heavily on hue, leading to the inclusion of pixels of debris and artifacts. As the w_S increased, the more detailed apertures on the grain surface were lost since they were lighter in saturation. Based on expert experience, weight values were heuristically chosen as 0.4 for w_H and 0.6 for w_S in order to produce the most consistent segmentation. To automate the selection of channel-merging weights, a training dataset of images with user-defined segmentation is needed to tune these two parameters. A simulated annealing

(SA) algorithm [56] can be implemented for automatic parameter selection [57]. This was not done in this study due to limited sample size, but could be implemented with a larger image training dataset.

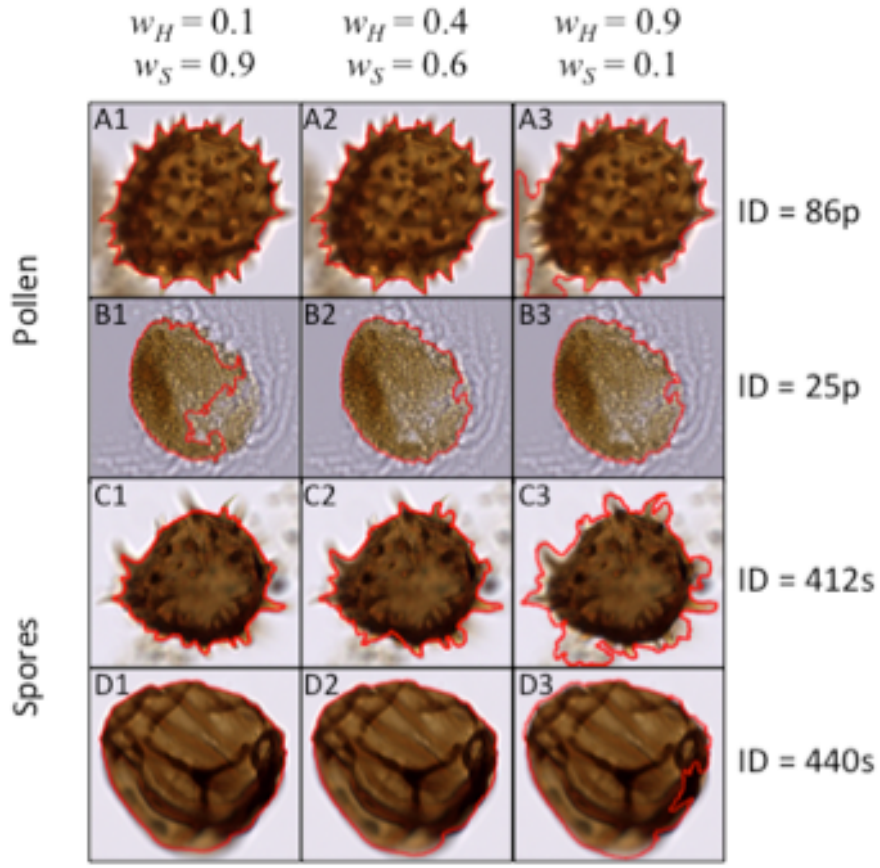


Figure 4.2 Weight configuration examples using two pollen grain images (row 1, ID = 86p; row 2, ID = 25p) and two spore grain images (row 3, ID = 412s; row 4. ID = 440s). Three segmentation results (highlighted red contours superposed on original grain images) are shown per each image example using different weight configurations for hue (w_H) and saturation (w_S) channels.

Next, the intermediate image was binarized using Otsu thresholding, which automatically selected a threshold value for binarization [8]. Morphological operations (non-linear operations related to the shape or morphology characteristics in an image) such as erosion, dilation, opening, and closing [3] were performed to separate the main body of the grains from any

debris or trivial particles that were not of interest in the analysis. Connected components [3] were identified to represent object candidates and only the largest one (presumably the grain) was kept. Finally, a Watersheding algorithm [58] was used to separate any remaining particles that were still connected to the main body of the grain. It is also possible to separate the grain from debris using the combination of weights described in the previous paragraph when there was a distinct boundary between grain and overlapped debris based on differences in pixel value of saturation, hue, and intensity or in surface texture. Segmentation is still a largely unresolved problem in image analysis research. It is widely recognized in image segmentation that when target objects overlap with debris, their boundary is blurred and undistinguishable, and segmentation performance has less consistency and accuracy. Human delineation may ultimately be needed to construct a reliable training set for our computer vision program to learn to separate objects from background. However, most of the grain samples in this study minimally overlapped with debris and efforts were made to confirm accurate segmentation.

4.2.4 Visual Feature Extraction

Once the pollen or spore grain is segmented, 69 visual features related to global visual characteristics (such as color, pixel value histograms, and textural patterns, listed in Table 4.2) and object morphology (such as convexity of convex hull, curvature of contour, and aspect ratio of bounding box, illustrated in Figure 4.3) are extracted from each of the four channels representing the original image:

hue, saturation, value, and gray scale. This produced a 276-dimension feature space in which individual images were placed.

Table 4.2 Visual features extracted from four single channel images.

Name	Description	#Features	Index
Threshold	OTSU threshold	1	color
Mean	Mean pixel value	1	color
STD	Standard deviation of pixel value	1	color
Histogram	1-dimensional histogram with 16 bins	16	color
Size	Grain object size	1	shape
HU	HU shape descriptors [15]	7	shape
Aspect ratio	Ratio of long edge to short edge of bounding box (Figure 4.3C)	1	shape
Compactness	*see Appendix*	1	shape
Convexity	*see Appendix*	1	shape
Form factor	*see Appendix*	1	shape
Roundness	*see Appendix*	1	shape
Solidity	*see Appendix*	1	shape
Perimeter	*see Appendix*	1	shape
Texture	Seven Haralick textures with five step sizes [4]	35	texture
		Total = 69	

Note: the numbers in the last column indicates the value per single channel image. All features are calculated within segmented grain objects only. Refer to Appendix for detailed calculation.

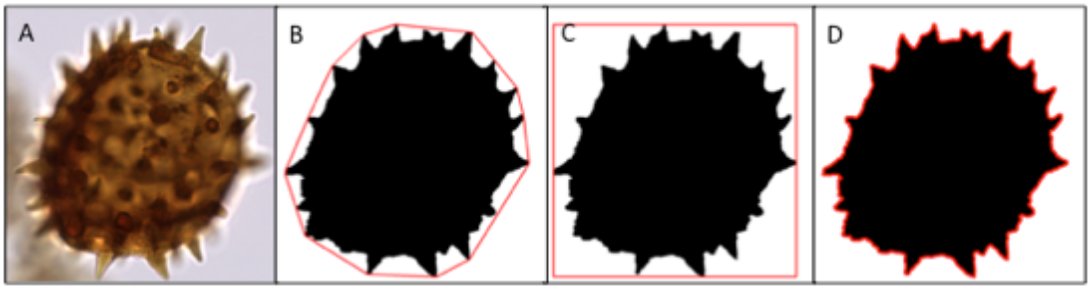
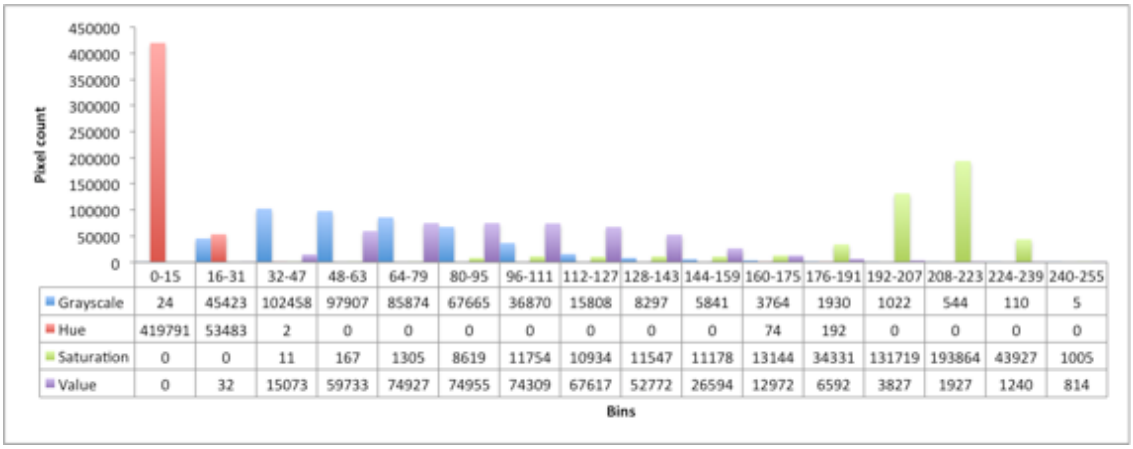


Figure 4.3 Image examples of selected features listed in Table 2. A: original image, B: Convex hull that encloses binarized pollen grain, C: Bounding box that encloses binarized pollen grain, and D: Contour that traces along the boundary of binarized pollen grain.

4.2.5 Morphology Content and Semantic Modeling

Palynologists use common qualitative terminology to describe and compare the morphology of pollen and spores [59]. However, complex

morphological features that are relatively easy for human experts to detect and describe linguistically are much more challenging for the computer to recognize numerically. To mimic the complex human process of identifying visual patterns, low-level visual features were extracted to represent the visual content in images. Examples of low-level features include: single channel histograms (Figure 4.4) Hu shape momentum descriptors [15], texture [4], etc.



	0-15	16-31	32-47	48-63	64-79	80-95	96-111	112-127	128-143	144-159	160-175	176-191	192-207	208-223	224-239	240-255
Grayscale	24	45423	102458	97907	85874	67665	36870	15808	8297	5841	3764	1930	1022	544	110	5
Hue	419791	53483	2	0	0	0	0	0	0	0	74	192	0	0	0	0
Saturation	0	0	11	167	1305	8619	11754	10934	11547	11178	13144	34331	131719	193864	43927	1005
Value	0	32	15073	59733	74927	74955	74309	67617	52772	26594	12972	6592	3827	1927	1240	814

Figure 4.4 Example histogram of an example image in four individual channels: grayscale, hue, saturation, and value.

In some image analysis research domains, such visual patterns are interpreted using high-level abstractions, called *semantics*. The extracted features can describe, to a limited extent, the visual content of grains, but are still not easily interpreted by the human analyst. This is known as the semantic gap [20]. To minimize the semantic gap, mathematical models are constructed using low-level features to map images to high-level trait semantics based on degrees of relevance. Using the mathematical formulas detailed in [29] and [60], each semantic representation is constructed as an association model using the concept of Possibilistic C-Means Algorithm [61] based on low-level visual features. This process is called Semantic Modeling (SM).

In this study, there are three morphology semantic categories for pollen images and three for spores, each of which consists of several exclusive semantic labels (Table 4.3). Using semantic modeling, each semantic label was represented as a semantic model of low-level visual features. Semantic model $M_\zeta$is trained based on a training dataset of images all labeled with semantic ζ. A trained semantic model M_ζ returns a relevance score for each database image for this specific semantic label. To reduce over fitting issues during semantic modeling and to estimate how well these trained models handle images that lack certain semantic labels, 10-fold cross-validation [62] was conducted in this study. In our study, an image was first represented by a multi-dimensional feature vector, which was then fed into each semantic model to calculate its relevance scores. These relevance scores were then used for automatic semantic annotation and semantic-based image retrieval.

Within each category, the higher the relevance score, the larger the possibility that an image has this particular morphology semantic. The model that produces the highest score in each semantic category determines the assignment of semantic label to an image. In this study, a grain image can be annotated with three semantic labels, each from a different category. For example, the relevance scores for a spore image are {(*pyramidal* =0.661, *plane-convex*=0.506, *reniform*=0.333) $^{lateral\ view\ shape}$, (*elliptic*=0.333, *circular*= 0.921) $^{polar\ view}$, (*radial*=0.921, *bilateral*=0.333) ^{symmetry}}.

Table 4.3 Semantic labels used to describe morphology of pollen and spore grains.

	Semantic Category	Semantic Label	# images
Pollen	Equatorial view	Prolate	50
		Spherical	138
		Oblate	25
		Unlabeled	184
	Polar view	Elliptic	137
		Circular	140
		Unlabeled	120
	Symmetry	Radial	172
		Bilateral	136
		Unlabeled	89
Spore	Lateral view	Pyramidal	52
		Plane-convex	28
		Reniform	23
		Unlabeled	25
	Polar view	Elliptic	23
		Circular	80
		Unlabeled	25
	Symmetry	Radial	80
		Bilateral	23
		Unlabeled	25

In Category "*lateral view shape*", *pyramidal* has the highest score compared to *plane-convex* and *reniform*. Therefore this image can be annotated as having a *pyramidal* shape in lateral view. With the same strategy, this image can also be labeled as *circular* in Category "*polar view*" and *radial* in Category "*symmetry*". In general, newly acquired images are not labeled. Relevance scores provided by semantic models will be useful for automatic annotation of images with undetermined semantics. Once the models are trained, no human intervention is needed for model selection and image annotation. Confusion matrices were used to visualize annotation performance for individual semantic labels. Average accuracy were calculated for pollen and spore samples in this study.

When trained semantic models are used for automatic semantic annotation, they are evaluated by annotation accuracy. After 10-fold cross-validation, the annotation accuracy is shown in the form of confusion matrix for pollen and spore morphology semantics (Table 4.4 and Table 4.5). In a confusion matrix, the value x in a cell (ζ, τ) means that x images with human-annotated semantic label ζ are annotated by computer with semantic label τ. Cell values are meaningful only when ζ and τ are from the same category. In an ideal scenario, we expect all images be annotated with correct semantic labels in each category. Therefore the confusion matrix should only have non-zero values in cells on the diagonal where row label (human-annotated semantic) and column label (computer-annotated semantic) are the same. In reality, errors cannot be completely eliminated in automatic annotation. For example in Table 4.4, among the 25 images that were labeled as *spherical* in *equatorial view shape* category, 18 images were annotated by computer correctly while the other 7 images were annotated as *prolate*. Then the accuracy of annotation for semantic label *oblate* is 18/25 = 72.0%. The average accuracy is 83.9% for pollen semantic annotation and 98.6% for spores.

Table 4.4 Confusion matrix of pollen image trait semantic assignment.

	p	***s***	***o***	***e***	***c***	***r***	***b***	**Accuracy (%)**
p	50	0	0					100
s	45	90	3					65.2
o	7	0	18					72.0
e				133	4			97.1
c				23	117			83.6
r						129	43	75.0
b						8	128	94.1

p=prolate, s=spherical, o=oblate, e=elliptic, c=circular, r=radial, b=bilateral

Table 4.5 Confusion matrix of spore image trait semantic assignment.

	p	***v***	***r***	***e***	***c***	***r***	***b***	**Accuracy (%)**
p	47	5	0					90.4
v	0	28	0					100
r	0	0	23					100
e				23	0			100
c				0	80			100
r						80	0	100
b						0	23	100

p=pyramidal, v=plane-convex, r=reniform, e=elliptic, c=circular, r=radial, b=bilateral

Queries using pollen images had an average MAP score of 0.81/1.00 and queries using spore images had an average MAP score of 0.93/1.00 (Table 4.6). The average search time for pollen and spore images was less than 0.2 second and as short as 65 milliseconds.

Table 4.6 MAP scores for semantic models trained over 10-folder cross-validation.

	Semantic Category	**Semantic Label**	**MAP**
Pollen	Equatorial view	Prolate	0.86
		Spherical	0.88
		Oblate	0.73
	Polar view	Elliptic	0.83
		Circular	0.70
	Symmetry	Radial	0.86
		Bilateral	0.79
Spore	Lateral view	Pyramidal	0.98
		Plane-convex	0.84
		Reniform	0.94
	Polar view	Elliptic	0.88
		Circular	1.00
	Symmetry	Radial	1.00
		Bilateral	0.88

4.3 Hierarchical Structure in Whole-Slide Pathology Images

4.3.1 Background

The cognitive burden of analyzing various types of histopathological patterns is rapidly increasing due to the application of an ever-expanding number

of new immunohistochemical biomarkers, which can shed light on the molecular machinery of underlying biological processes in a diseased tissue. It is likely that some subtle, or even obvious, visual patterns can be overlooked during evaluation. However, it is expected that the acceptance of Whole Slide Imaging (WSI) technology in routine pathology practice will allow a computerized analysis of tissue sections for diagnostic purposes. Computer vision techniques may help to reduce the chances of overlooking visual patterns. Several studies [63, 64] on automating histopathological image analysis focus mainly on identifying pathological objects that are relevant for diagnosis with certain degrees of success. Yet, there is additional rich diagnostic information, such as anatomical structures, their spatial relationships, and finer pattern underlying those structures, which also need more attention and extensive studies. However, due to the huge file size associated with WSI, it is not trivial to develop algorithms to support such studies. To address these issues, we have developed a computational approach to detect follicles, one type of anatomical structure, in immunohistochemical (IHC) stained slides and then further measure protein expression, geometry, and spatial information to support diagnosis in a 'coarse-to-fine', multi-resolution fashion. This is expected to reduce the consumption of processing time and potentially reduce inter- and intra-pathologist variability in diagnosis.

4.3.2 Follicle Detection

We have collected whole slide IHC images of FL cases and reactive hyperplasia cases using an Aperio® ScanScope® CS digital scanner [65].

Pathologists diagnosed all the images. The images are organized into multi-resolution Gaussian pyramids which can represent visual patterns of objects at different scales, i.e. from coarse to fine layers are tissues, follicles, follicular structures, cells and sub cellular structures. As resolution gets higher, richer pixel information will be retained and image size gets bigger at the same time that will result in the cost of a much longer time for image processing.

To take the advantage of the characteristics of Gaussian pyramid structure in our image analysis processes, we focus on specific image analysis at each layer and narrow down the processing tasks as we go down to higher resolution layers. At each layer, images are further divided into small tiles from which visual features are extracted to represent informational content of individual tiles. Various types of visual features can be extracted based on the tasks pursued on specific layers. Features include pixel values, grayscale histogram, and co-occurrence textures [4]. As regard to pixel values, we first convert the image from RGB (red, green, and blue) color space into HSV (hue, saturation, and value) color space, which is expected to be capable of separating the color information from intensity and is more practical for human interpretation [3]. In our collection of slides, pixels from different anatomical structures, such follicles and inter-follicular regions, have different hue ranges, i.e. follicles, including germinal centers and mantle zones, have brown hue values while pixels in inter-follicular regions have blue hue values (Figure 4.5). Next, we applied Gaussian filters to each channel image to blur out potential noises. In order to keep critical pixel information while eliminating noise, we empirically designed our filters window sizes according to the layer resolutions, i.e. the window size is 3x3 on layer with

down sample size of 32 (7.936 microns per pixel) and it is enlarged to 15x15 on layer with down sample size of 4 (0.992 microns per pixel).

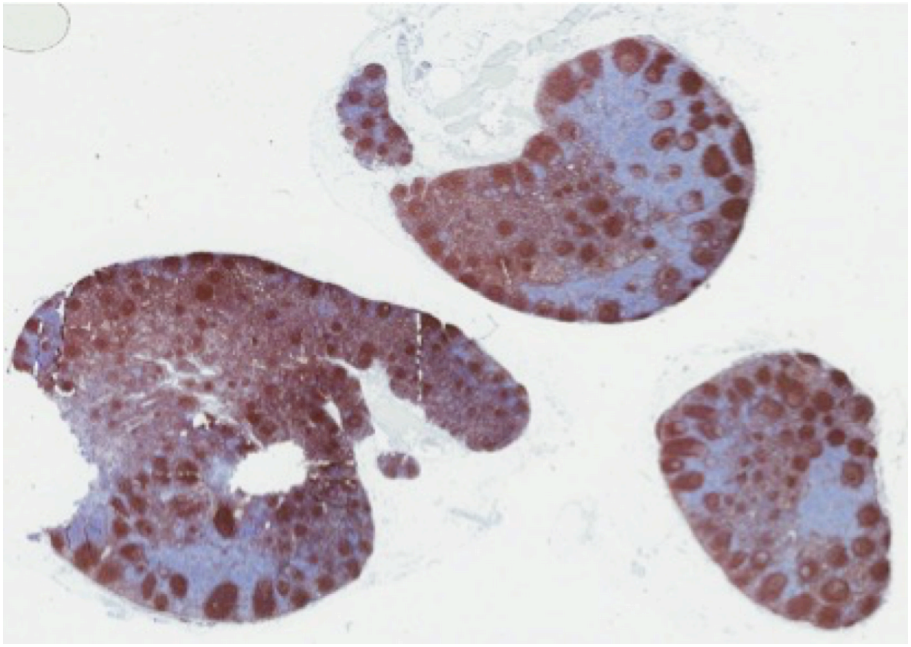


Figure 4.5 An IHC-stained whole-slide image (CD23) diagnosed as reactive hyperplasia.

Tiles on the top layer are grouped, using K-Means clustering algorithm (K = 3), into regions that correspond to three anatomical structures: germinal centers, mantle zone and inter-follicular regions. Then the clustering result is mapped onto finer layers that will provide detailed patterns, such as color intensity, shape, and texture, for each of the three anatomical structures. Clustering process is again applied on a finer layer with the *a priori* identification of structures so that finer segmentation adjustment is expected. By collecting the evidence of spatial relationships of follicular structures among these three anatomical structures, it becomes possible to compute measurements of pixel density and texture in germinal centers and mantle zones, ellipse fitness (ratio of

unfit are to fitted area), follicle size, and distribution of follicles in the tissue, as well as to measure the intensity of IHC staining of surface biomarkers.

4.3.3 Results

We selected two IHC-stained slides for our preliminary study. One slide is CD20-stained, Grade II follicular lymphoma and the other is CD23-stained, reactive hyperplasia. The preliminary results show that all the manually identified follicles were detected by automatic computer vision program. Moreover, we also measured the segmentation accuracy according to the similarity index defined in [66] as

$$S = 2n\{D \cap M\}/(n\{D\} + n\{M\}) \tag{4.2}$$

where D and M are objects of interest that were automatically detected by our algorithms and manually by the pathologist, respectively, and $n\{\cdot\}$ means the area, or number of pixels, of each object. The average segmentation accuracy achieves 83.09±6.25%. Figure 4.6 shows an example of the detection of anatomical structures, follicles in this case.

Detection of anatomical structures from multiple layers of resolution has a 'coarse-to-fine' process that utilizes the information given on each layer so that general structures are identified first as regions of interest and then finer objects of interest can be found within the segmented regions. This process will reduce the consumption of processing time at an early stage and dedicate more on finer processes. Automatic detection and measurement of follicles provide an easier and potentially more reliable way for pathologists to analyze WS IHC images and provide support for their diagnoses based on quantitative visual information

extracted from images. Moreover, it has great potential to reduce inter- and intra-pathologist variability in diagnosis. Furthermore, image analysis performed in a hierarchical fashion to identify macro and micro visual object patterns can significantly reduce the processing time and allow pathologists to concentrate their attention on more complex tasks in histopathological studies. With proper extension and improvement, we believe that our method can be instrumental in diagnosis of malignancies related to follicular structures.

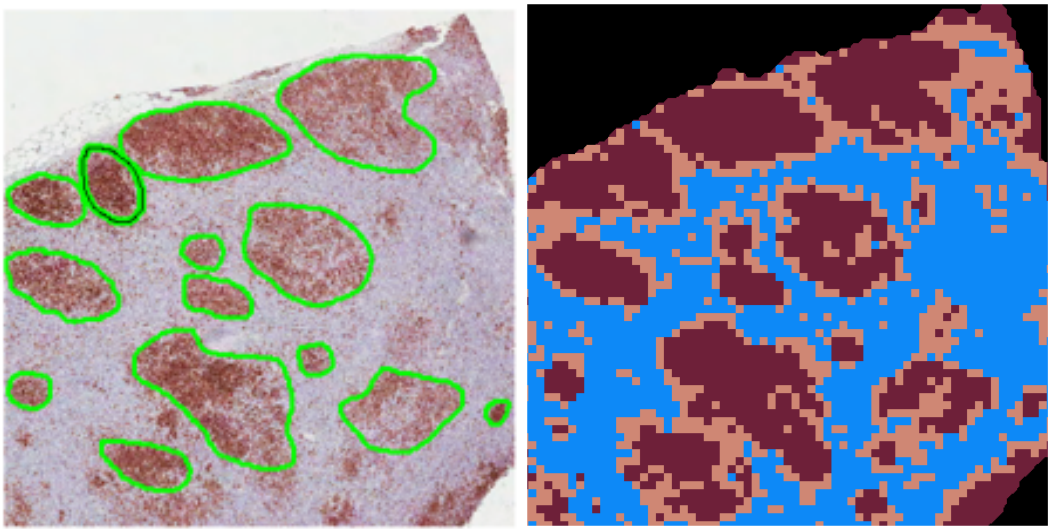


Figure 4.6 Detection of anatomical structure. Follicles (left, green lines) that are manually identified by pathologists are detected automatically by computer vision algorithms (right, dark brown germinal centers and light brown mantle zones).

4.4 Pathology-Bearing Regions in HRCT Images of Lung

4.4.1 Background

Domain experts, such as physicians and experienced users, always look for some distinct visual patterns appearing in the images that represent certain disease characteristics. These visual patterns can be generally put into groups and we refer to them as *perceptual categories* (PC) [67]. In this study, there are five perceptual categories represented in the image collection: emphysema (EMP),

cysts (CYS), ground-glass opacities (GGO), honeycombing (HON), and bronchial structures (BRO). Because it is not uncommon to observe different visual patterns, belonging to different PCs, in the left and right lungs (Figure 4.7) the two lungs are analyzed separately.

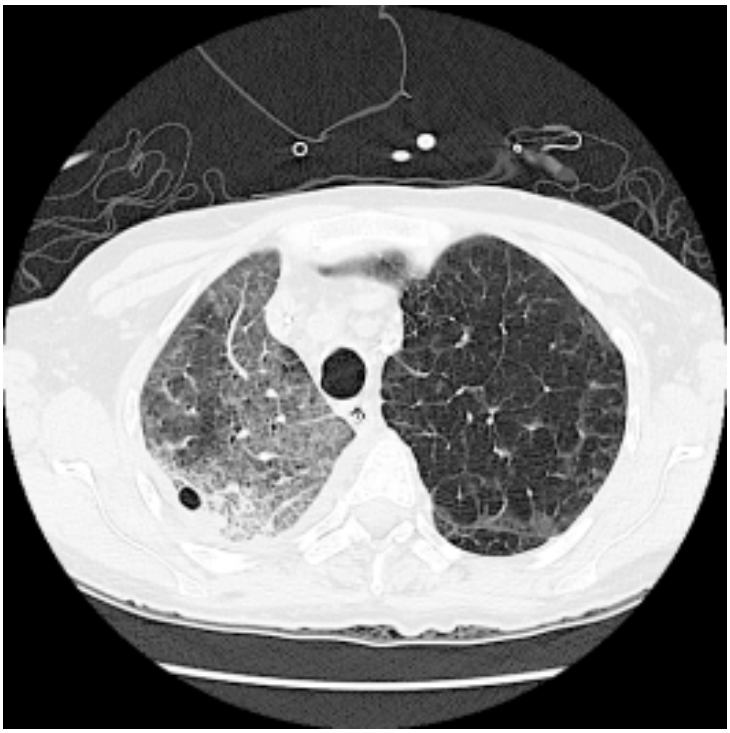


Figure 4.7 A HRCT lung image with a PC of ground-glass opacity (GGO) in the right lung (the left side) and an emphysema (EMP) perceptual category in the left lung (the right side).

As part of a CBIR system development for HRCT images of lung, we need to extract visual patterns that are relevant to these aforementioned PCs. From our experience, to build a successful CBIR system with medical imagery, the following assumptions must hold: (1) medically meaningful objects of interest can be identified; (2) the selected features are indeed sufficient to characterize the various appearances in the image set; and (3) feature values are properly extracted and accurately represent the true patterns residing in the images. These

sound fairly basic, but are difficult to ensure all the time in practice. A significant amount of this difficulty lies in adjusting and adapting all the various parameters used in the object segmentation and feature extraction algorithms. This problem is common to most of the applications using CBIR technique and is therefore a very valuable issue to be studied. The approach presented in this section makes parameter tuning automatic according to provided medical images instead of the developer's empirical settings.

4.4.2 Modularized PC Recognizers

After segmenting the right and left lungs from the background of the image, we can analyze each lung independently for various patterns using customized algorithms called *modularized PC recognizers,* or *modules* in short. In each module, different image processing algorithms and filtering criteria, such as grey scale thresholding, connected components, topological characteristics, spatial relationship information, etc., are designed to filter out artifacts and at the end of each module objects of interest representing this PC's patterns are left to calculate low-level features.

Each step in each module utilizes thresholds, decision criteria, or some other logic to make decisions about whether segmented objects are indeed useful or merely represent artifacts. All these steps ultimately work together and collectively lead to a final image segmentation result, the quality of which can drastically affect feature accuracy. Usually assigning parameter values will greatly rely on domain experts' knowledge as well as researchers' accumulated experience through the development of the system. The selection of the

configuration of various parameters can be considered as a combinatorial problem. Because a change in the value of one parameter will likely affect the results of other steps in the algorithm, it is almost impossible to configure the entire array of parameters manually.

As an example, the *cystic structure (CYS) module* contains seven parameters that need to be computationally optimized. Ideally, an image with cystic structures that is analyzed by the CYS module should be segmented into as many cysts as possible, while an image with only emphysema that passes through the CYS module should result in no segmented objects. If the parameters in the CYS module are not properly set, using empirically configured default values for example, then it is quite likely that some low attenuation regions may also be extracted incorrectly and using features calculated from them the image may be considered as CYS. The hope is that by tuning parameters, this insufficiency can be greatly reduced.

Figure 4.8 demonstrates comparisons of each module's resulting images when applying ideally best parameter configurations; default configurations based on empirical experience; and computationally optimized configurations using simulated annealing. In the next section, we will explain our novel methods to automate the selection of crucial parameters and ultimately improve visual content and image retrieval of HRCT images of lung.

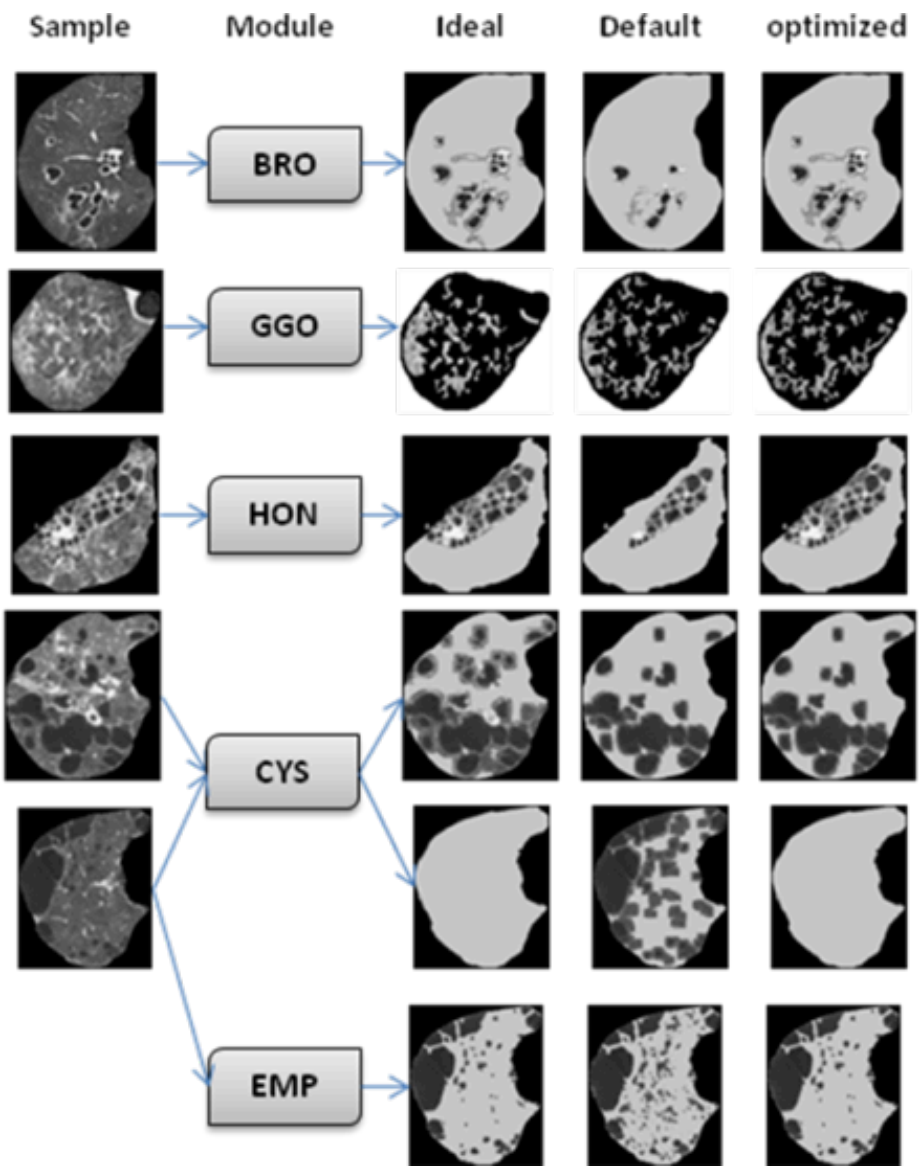


Figure 4.8 Modularized PC Recognizers and extracted results from various parameter settings.

4.4.3 Improve Visual Content Extraction with Automatic Parameter Tuning

Simulated annealing [56] is one of the global search methods used for optimization problems. It is useful to deal with combinatorial problems with finite but relative large feasible sets for parameters [68]. Random trials of possible values are tried, and the performance of each trial is evaluated using a customized cost function. A probability of accepting a "worse move" is used to

give the trial a chance of "climbing out of local minima" while searching the entire feasible set. In our study, we extend the method of simulated annealing to accommodate the needs to adjust parameters from each module. Parameters in each module are computationally optimized one by one in an order based on (1) their positions in the module as well as (2) their perceived influence on the final results. For the sake of reasonable computing time, we optimize the parameters one at a time. The configuration is updated after finishing each parameter's SA tuning procedure. Empirically derived values are used if the parameters have not been optimized. Secondly, after optimizing one parameter, the temperature is "reheated" to the initial temperature T_0. This gives the next parameter tuning step an equal chance to move around and escape the local minima obtained from previous parameters' SA results.

4.4.3.1 Overall Process of Automatic Parameter Tuning

The system operates as follows (Figure 4.9):

- *Initial Settings:* Specify working module, order of parameters to be tuned, initial temperature T_0 and stopping criteria which include maximum iteration number, frozen temperature T_f, control value c of exponential cooling schedule11 and target minimum cost value12.
- *Select Parameter*: In the module's configuration parameter list, pick one parameter according to predefined order.
- *Simulated Annealing*: Use simulated annealing to find the optimal value of picked parameter.

- *Update Parameter*: Set the value of this parameter to the optimal value found in the SA step and proceed to tuning the next parameter.
- *Final Configuration*: All the tuned parameters together form the final configuration for the working module. Once these are all determined, proceed to the next module

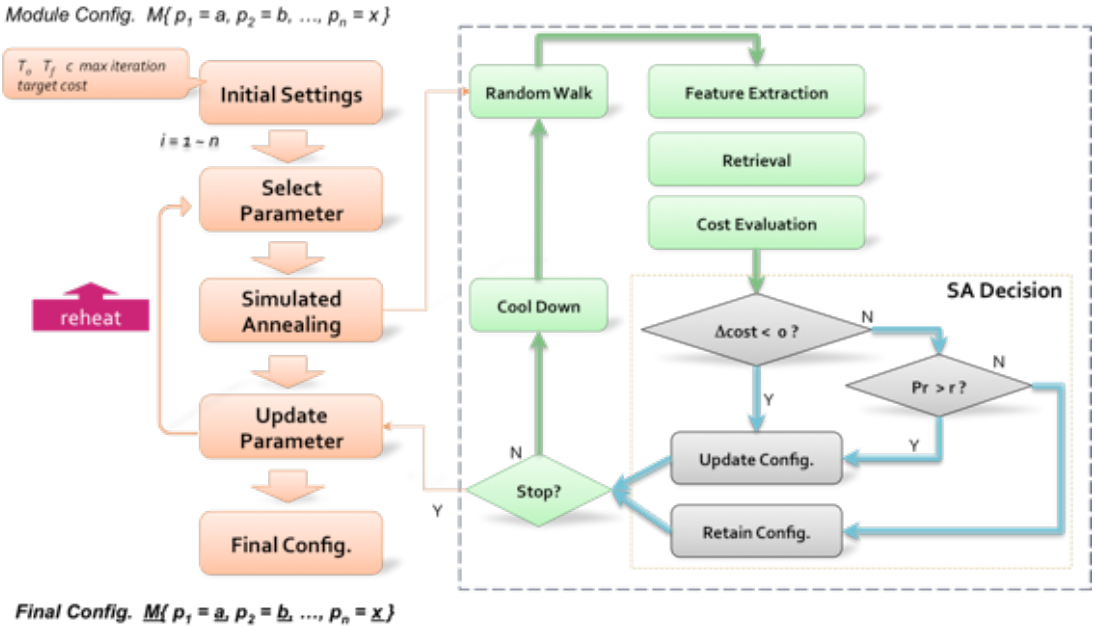


Figure 4.9 Overall process of automatic parameter tuning with SA.

4.4.3.2 The Simulated Annealing Step

The procedure for performing simulated annealing, shown in the big dashed box on the right side of Figure 3, for one parameter in a module includes the following steps:

- **Random Walk**: Adjust the value of the parameter by one small step. Step size and valid walking range are predetermined for each parameter using empirical knowledge. Use this new value keeping other parameters fixed to form a new configuration.

$$value_i^j = value_i^b + step_i^j,\ step_i^j = 2 * step_i^0 * r - step_i^0 \tag{4.3}$$

$$value_i^j \in [\min value_i, \max value_i]$$
$$step_i^j \in [-step_i^0, step_i^0]$$

where $i = 1,2,\cdots,m \quad j = 1,2,\cdots,\max iteration$

$value_i^j$: j^{th} trial for parameter i in a module
$value_i^0$: initial value for parameter i
$value_i^b$: current best value of parameter i
$step_i^j$: step length in j^{th} trial for parameter i
$step_i^0$: maximum step length for parameter i
r : uniform random number between 0 and 1

- **Feature Extraction**: Use this new configuration to segment out objects of interest and extract low-level features for all images.
- **Retrieval**: At each annealing step randomly select a set of images that are labeled as the current module's PC to search against remaining images and retrieve top ranked results. This random selection of query images helps to limit the effect of over-fitting. In our study, we query 10 times and pick top 30 results for each query.
- **Cost Evaluation**: Cost functions are derived using two methods, average precision at seen relevant documents and F_β measure. Images with same labels as query image are considered as relevant results. Averaged cost over all random query evaluations are used as each annealing step's cost value.

$$\text{Cost} = 1 - \text{Ave.Prec} \quad \text{or} \quad \text{Cost} = 1 - \text{F}_\beta \tag{4.4}$$

$$\text{Ave.Prec.} = \frac{1}{\text{k}} \sum_{r=1}^{k} p(r) \tag{4.5}$$

k : # relevant result documents
$p(r)$: precision at rth relevant document

$$\text{F}_\beta = \frac{(1+\beta^2)\, precision * recall}{\beta^2 * precision + recall} \tag{4.6}$$

$precision$ = # relevant / # total results
$recall$ = # relevent results / # total relevant

- **SA Decision**: Make a decision on whether to accept the new move. If the cost from the new configuration is less than the current best configuration, set this configuration as the new current best. If not, calculate the probability of accepting a "worse step" and accept it if this probability is greater than a uniform random number; otherwise we retain the current best configuration.
- **Cool Down**: Reduce the probability of accepting a worse step by cooling down the system temperature. The SA step stops when one of the following three stopping criteria occurs: steps reach the maximum, temperature drops to a frozen state, or cost reaches to the predetermined low value.

In order to perform annealing in a controlled manner, a cooling schedule and other stopping criteria are set. Raittinen & Kaski suggest that the initial and frozen temperatures can be the maximum and minimum cost differences, respectively [69]. Since the cost function in our study ranges in [0, 1], the initial temperature is set to be 1 and the frozen temperature is set as 0.001. The control value of exponential cooling schedule is 0.98 and target low cost is 0.05 or equivalent to say the target precision of retrieval is 95% for the method of average precision at seen relevant documents or the target F_β score is 0.95 for the method of F_β measure. The maximum number of iterations for each SA step is set at 300 in order to make the computing time reasonable.

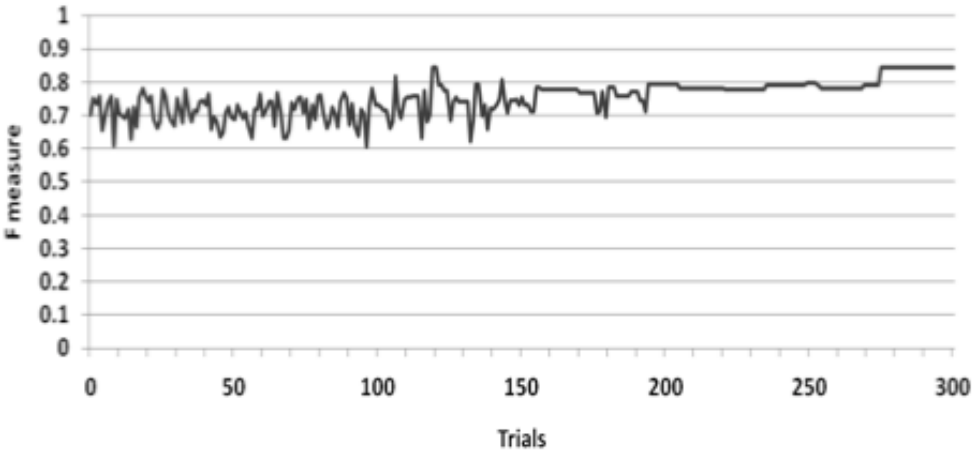


Figure 4.10 Precision of each step's saved best configuration for parameter #3 in CYS module over 300 trials.

Figure 4.10 shows an example of optimizing CYS module's third parameter. The F_β measure with $\beta = 0.33$ increases gradually as temperature drops down and becomes stable around 85% after 300 trials. We use $\beta = 0.33$ to emphasize 3 times more of precision than recall due to the intension that users would pay more attention to see top ranked results to be as many relevant results as possible rather than browse all possible results.

4.4.3.3 Parameter Tuning Performance

We selected 303 HRCT lung images, in which a total of 394 left or right lungs are labeled individually and with a consensus of two radiologists according to their visual patterns with five perceptual categories. Images are all gray scale images with dimensions of 512 x 512 pixels.

In our study, each module has its own parameters that control performance quality, and the scenarios for these modules are different due to varying characteristics of each perceptual category's visual patterns. Therefore, we apply a "divide-and-conquer" approach to optimize parameter configurations

for each module. Details regarding the distribution of perceptual categories, parameters and low-level features extracted from images are listed in Table 4.7.

Table 4.7 Distribution of images, parameters, and features. There are 47 global features including statistics of overall gray scale (12) and textural measurements (35).

Module	Lung No.	Parameter No.	Feature No.
BRO	62	5	7
CYS	52	7	8
EMP	121	6	6
GGO	106	6	6
HON	53	5	5
Total	394	29	69

Using simulated annealing to find the computationally optimal configuration shows improvement of retrieval's effectiveness for all modules using both cost evaluation methods (Figure 4.11). The average retrieval precision improves from 79.49% to 93.73% which is a relative 18.021% increase using average precision at seen relevant documents, and it increased relatively 28.93% which is from 53.52% to 68.85% using F_β measure. One-tailed student's t-tests are used to determine the statistical significance of improvement (Table 4.8). The null hypothesis is "mean effectiveness (precision or F_β score) is not different after SA optimization" and the alternative is "SA optimization improves the mean effectiveness". The p-values using two methods are 0.55±0.001 (precision only) and 0.058±0.001 (F_β).

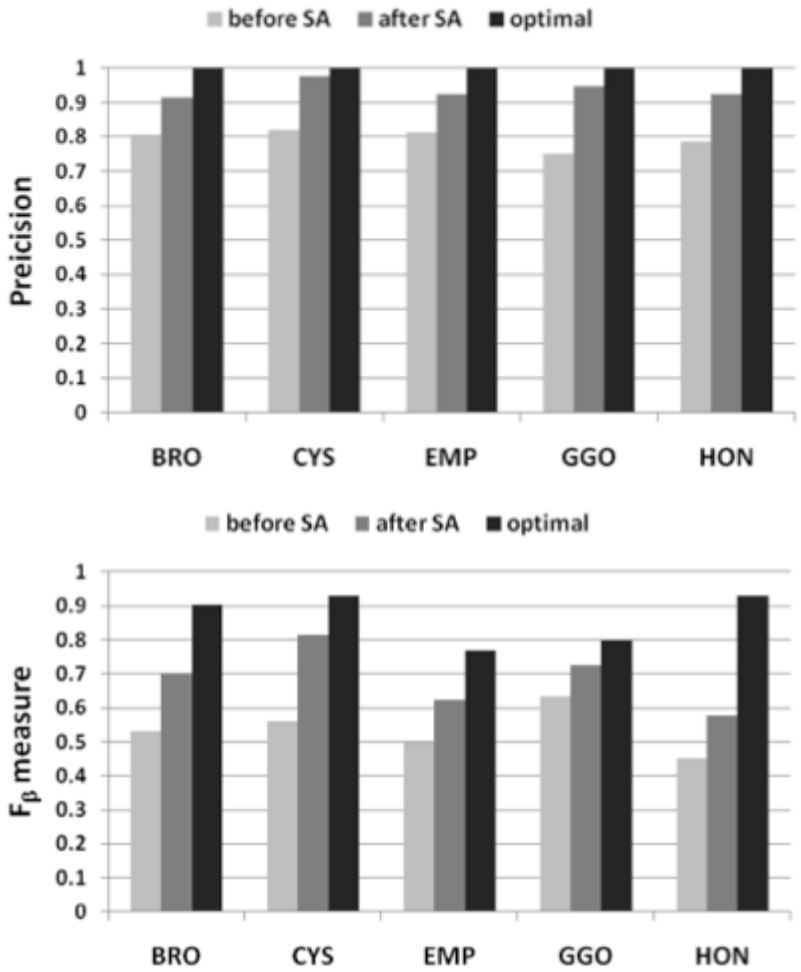


Figure 4.11 Effectiveness increase before and after optimization using average precision at seen relevant documents (top) and F_β measure (bottom).

Table 4.8 Comparison of mean and variance of precision/F_β measure before and after SA for all five modules.

		BRO	CYS	EMP	GGO	HON
Pre	Before	0.81±0.069	0.82±0.099	0.81±0.037	0.75±0.069	0.79±0.049
	After	0.92±0.002	0.98±0.001	0.93±0.005	0.95±0.005	0.92±0.002
F_β	Before	0.53±0.077	0.56±0.114	0.50±0.029	0.63±0.025	0.45±0.076
	After	0.70±0.054	0.81±0.010	0.62±0.023	0.73±0.006	0.58±0.005

4.5 Summary

In this chapter, we demonstrate our research works on visual content extraction for biological and medical images. Particularly, the visual categories that describe the morphology characteristics of Neotropical pollen and spore

grains were discovered using low-level visual features as well as high-level semantic modeling, preparing us for content-based image retrieval of grain images with query image examples (Chapter 5). Follicles in whole-slide IHC images were identified using hierarchical "top-down" method, preparing the processed image for extracting pathological meaningful content and eventual help us for visual category discovery and computer-assisted diagnoses (Chapter 6).

When a group of computer vision algorithms are used to identify objects of interest in an image, parameters in these algorithms need to be treated carefully. Empirically derived values may be insufficient for identifying objects and therefore hurt the quality of features. The customized version of simulated annealing was adopted in our research works for HRCT image of lung and helps to improve the retrieval precision for all modules. This approach can be applied to other image and information retrieval problems that need to deal with combinatorial problems and do not have a direct and analytic cost measurement.

In the next chapter, we will emphasize how these extracted visual content benefit the performance of content-based image retrieval.

CHAPTER FIVE

CONTENT-BASED MEDICAL AND BIOLOGICAL IMAGE RETRIEVAL

In previous chapters, we have introduced (1) several web-based image management systems for biological and medical images and (2) various applications on visual content extraction. They can be regarded as a front-end interacting with users and a back-end mechanism for translating raw images into computer-understandable visual features. In this chapter, we will present the last aspect that bridges these two ends together, delivering a functioning content-based image retrieval system.

5.1 Introduction

Content-based image retrieval (CBIR) is a technique for retrieving images based on image content using extracted features, such as color, texture, and shape [70]. When textual descriptions and annotations are limited or not available and needs more understanding and development, analyzing the content of images becomes more powerful. Moreover, patterns that are not easy for humans to pick up on can be detected automatically by computer vision techniques, and this information can contribute to even better accuracy in capturing critical patterns for diagnoses. The typical steps to performing CBIR include identifying the important characteristics in a set of images, designing computer algorithms to directly or indirectly measure these features, and finally utilizing existing or developed indexing structures for fast and efficient retrieval of visually similar images. This technique first appeared in the mid 1990s with the development of systems like QBIC [71], PhotoBook [72], and VisualSEEK

[73]. Over the past decade, the field has steadily grown and matured, and these techniques have since been applied to a wide variety of applications, including geospatial intelligence [74], astronomy [75], and protein structure comparison [76]. Another relevant collection of CBIR applications have been in the field of medical informatics [22].

5.2 Multi-Module CBIR System of HRCT Images of Lung

Medical images play a critical role in many aspects of clinical routines such as radiology, neurology, pathology, endoscopy, cardiology, dermatology, etc. Clinical professionals refer to medical images to examine diseases and make diagnoses according to the visual patterns observed in the images in conjunction with other medical data and observations. This results in a huge amount of medical imagery from various modalities being generated daily and needing to be reviewed, compared for diagnoses, and then archived in systems like picture archiving and communication system (PACS) for future reference and also for medical training purposes. According to [77], an estimated 62 million scans were collected in 2006 in the United States alone, compared to about 3 million scans in 1980. Thus, it is almost not feasible for radiologists to go through all the scans and potentially dig out previous similar cases. Therefore, turning to modern technologies, such as computer vision, database management, and information retrieval, is promising to ease this burden and may even improve diagnoses by backing up the findings with previously diagnosed similar cases.

Unlike images in general purpose CBIR systems, medical images usually do not have rich color information, and users primarily pay attention to

pathology-bearing regions, which can be difficult to extract as they are oftentimes not that distinguishable from the surrounding context. From our experience, to build a successful CBIR system with medical imagery, the following assumptions must hold: (1) medically meaningful objects of interest can be identified; (2) the selected features are indeed sufficient to characterize the various appearances in the image set; and (3) feature values are properly extracted and accurately represent the true patterns residing in the images. These sound fairly basic, but are difficult to ensure all the time in practice. A significant amount of this difficulty lies in adjusting and adapting all the various parameters used in the object segmentation and feature extraction algorithms. In section 4.4 we have introduced our research works in building modularized perceptual category recognizers for extracting visual content from HRCT images of lung. With a novel approach to automatically adjust multiple parameters in these modules, we are able to provide an improved CBIR system for radiologists to search for HRCT images that bear similar perceptual categories.

Figure 5.1 depicts a retrieval result using a Google-like query by image example method. The query image (top) is the left lung of an HRCT lung image that has bronchial structure patterns, which are dark lumen surrounded by thick walls. Using the features extracted from the query image, most similar images are retrieved and ranked based on overall similarity in the ranked order from top-left to bottom-right. We can see these images come from different patients with certain variation in appearance but all share similar visual pattern, BRO.

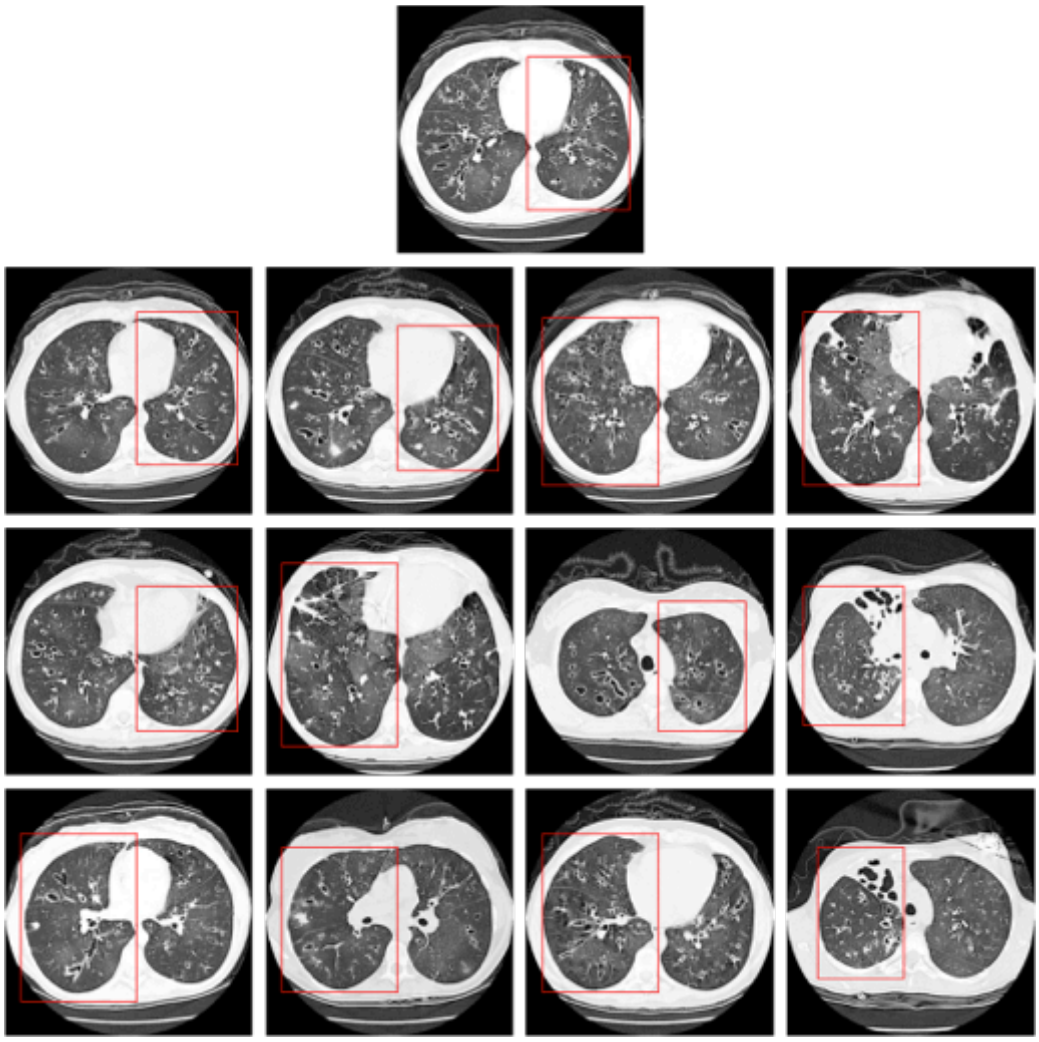


Figure 5.1 Results from the CBIR system for HRCT images of lung.

An accurate retrieval system in medical databases that archives medical images, clinical reports, laboratory data and other related information can help health care professionals and medical students search for similar medical cases and assist them in diagnoses. For example, pulling out top ranked CT scans that have most similar disease patterns will provide radiologists a means for differential diagnosis. Moreover, combining images with other clinical information and observations can provide necessary information for more accurate and efficient diagnoses.

For this CBIR system for images bearing multiple PCs, in order to examine all possible patterns, the retrieval results from each module need to be aggregated in a proper fashion, usually by treating features equally. However, this may not be the optimal solution for the reason that subsets of features may contribute more than others in the performance of retrieval. Therefore, we analyze the retrieval results in each module and generate a weighing scheme to make retrieval customized to individual search activity.

The procedure of multi-module retrieval involves two steps. First, visual features extracted from one module are used to retrieve top most similar images. These images may include images having patterns from same perceptual categories as the query image or from other perceptual categories. The entropy of each module's top 20 images is calculated as follows.

$$e_i = -\sum_{j=1}^{n} p_{ij} log p_{ij} \text{ where } p_{ij} = \frac{\#PC_{ij}}{\#Total_i} \tag{5.1}$$

$$w_i = 1 - e_k, \quad i, j = 1, 2, \dots, n \tag{5.2}$$

$$W = [w_1\ w_2\ \dots\ w_n]^T \tag{5.3}$$

Percentages of result images that belong to each PC are calculated to get the entropy of results. A more homogeneous result will result in a higher value of weight for that module in the weight vector. Next, the entire search results from all modules are unioned by applying the weight vector into a single pool of images. New similarity of each image in each module is calculated as

$$sim_{ik} = \frac{d_{ik} - \min\{d_{ik}\}}{\max\{d_{ik}\} - \min\{d_{ik}\}} * w_i \tag{5.4}$$

$$where\ i = 1, 2, \dots, 5; and\ k = 1, 2, \dots, n_i$$

The normalized Euclidean distance between the feature vectors of *k*-th image in results from module I and the query image is multiplied by the weight factor obtained from previous step. Top ranked images from this pool using new similarity scores will be the final retrieval results.

Aggregating search results from different modules based on particular perceptual categories' visual patterns based on search results' entropy helps to rearrange the similar cases. The preliminary study shows an increase of average precision from 69.5% using multi-module method to 76.6% be retrieval without weighing scheme.

5.3 Finding Similar Grains in Neotropical Pollen and Spore Images

As introduced in sections 3.2 and 4.2, I have participated in a long-term collaboration among palynologists, computer scientists and informaticians in an attempt to develop computational and informatic solutions to streamline the process of palynology analysis for efficient and reliable data management, analysis, and retrieval. To our knowledge, our work is the first attempt to develop a unique search engine that utilizes image-based morphological content for grain image retrievals in palynology. We report the following approaches.

First, we applied and extended a suite of image analysis algorithms and toolkits to automate the process of detecting grains from artifacts (debris and organic matter other than pollen and spores, common to fossil palynological slides) and calculated morphological features based on shape and texture. Next, association rule mining [78] was integrated into our methods to assist experts in trait annotation based on extracted features and a continuously updated expert

knowledge base. We then utilized information retrieval methods [79] to provide fast and accurate data management and image retrieval. The morphological features identified by our automated analysis were used to determine image semantics (abstract presentations of morphology) that formed the basis of novel tools for automatic semantic annotation, semantic-based image search, and content-based image retrieval by image examples.

5.3.1 Database Design for Multi-modal Information Integration

For an image database to be effectively used in taxonomic classification and customized image retrieval, accurate metadata are as important as novel search algorithms. Our database structure was designed to be flexible for database management across geographically remote sites and allows for sustainable growth over the time with the incorporation of new palynological images and the participation of new analysts. Instead of storing data in single files such as spreadsheets or printed catalogs, images and their metadata are stored in a relational database where the shared data structure and data relationships are carefully designed and maintained to avoid duplication or accidental modification. The entity relationship diagram (ERD) illustrates the database structure and its tables with relationships that ensure data integrity and handle dynamic data changes such as insertion, deletion, and update (Figure 5.2).

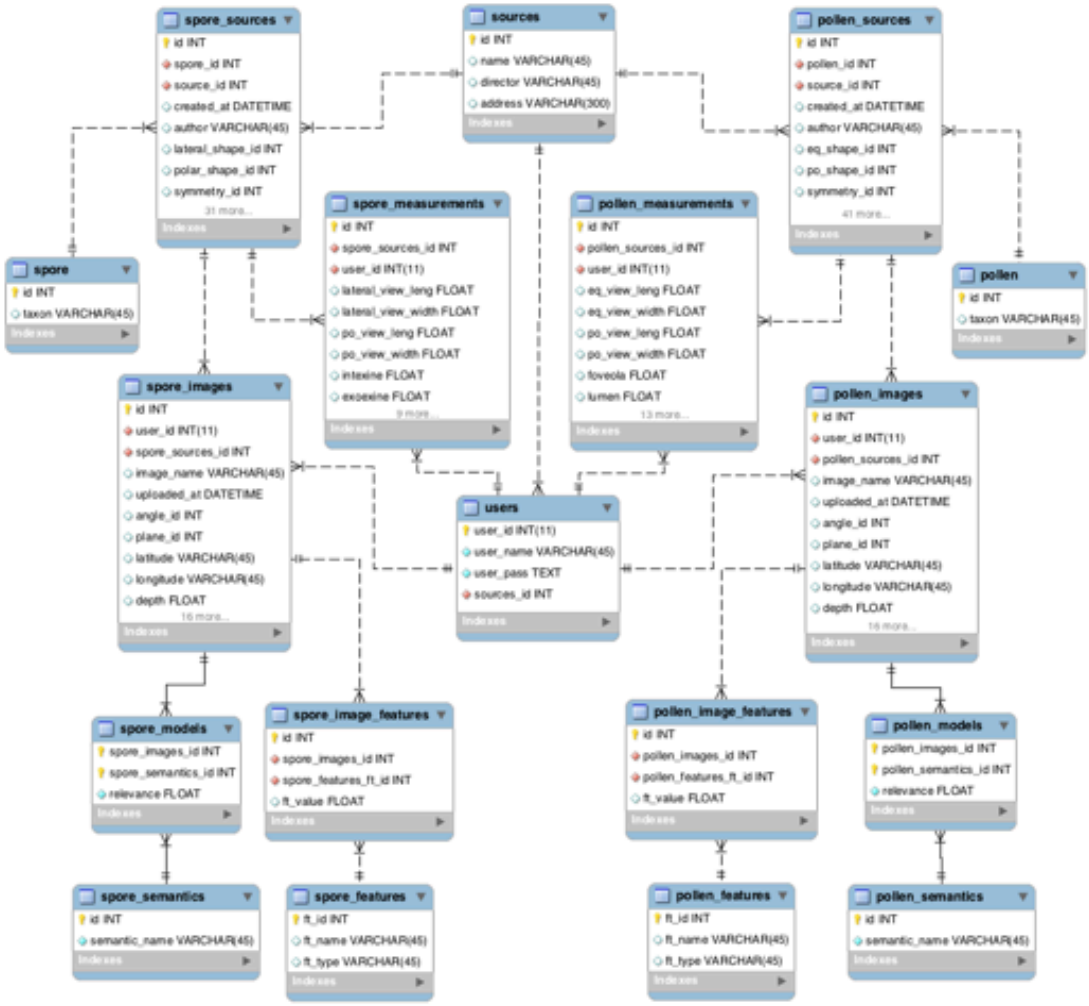


Figure 5.2 Entity Relation Diagram (ERD) of database design. Entities and their relationships are represented as tables with attributes and connected using crow's feet annotation. For example, the relationship between *pollen* and *pollen_images* is a one-to-many identifying relationship. Specifically, one pollen taxon can have multiple images and each record in *pollen_images* must reference to only one and only one record in *pollen*. There are four pollen-related tables on the right-hand side and four spore-related tables on the left-hand side. Tables from both sides share similar structure and reference to two common tables - users and sources. Note: There are 76 tables and over 200,000 records in the database. There are 49 attributes in table *pollen_sources* and 39 in table *spore_sources*. For simplicity, some auxiliary tables and secondary fields are omitted in this figure. Only the most relevant tables and fields are shown.

In the ERD, tables on the left sides are designed for spore taxa and pollen-related tables are on the right sides with same table structures. In addition, to

link multiple research groups for cross-site research, the *miocene_sources* table and the *miocene_users* table store information of research teams and palynologists who collected the images. With two relationship tables, *miocene_pollen_sources* and *miocene_spore_sources*, two sides are linked together to enforce relationship dependencies. With this database as the backend, a web-based system was built for palynologists to interact with stored data and search for grain images. The system provides not only text-based species search (Figure 5.3), but also image searches based on trait semantics (Figure 5.6) and visual content (Figure 5.8 and Figure 5.9) with personalized search criteria.

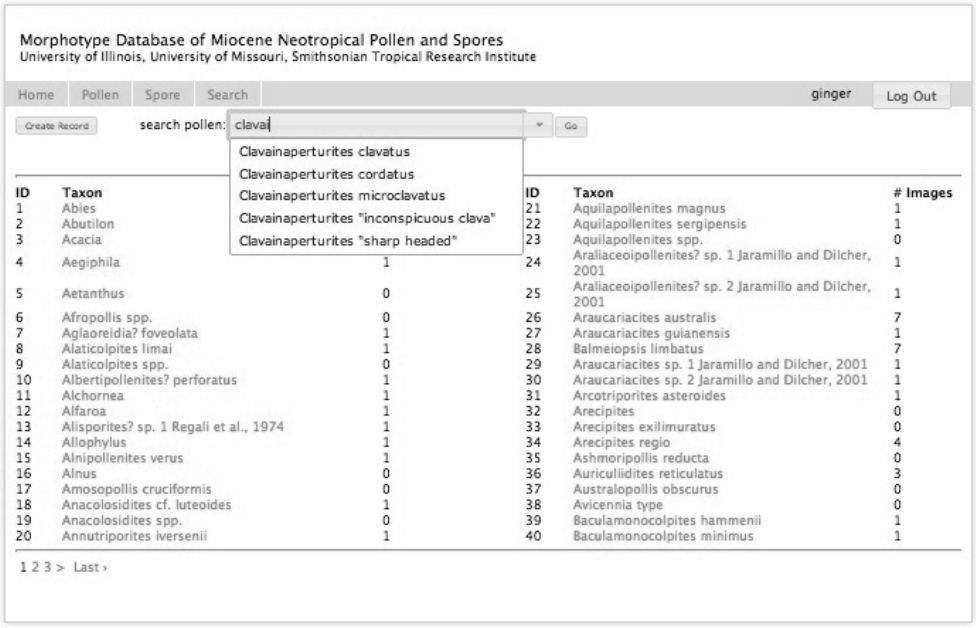


Figure 5.3 Searching for pollen taxa by name. All existing taxa in the database are listed on the webpages ordered by taxon ID. User can choose to search taxa by their scientific names by typing in the text field. Auto-complete hints help users to quickly narrow down the list.

5.3.2 Image Search using Semantic Models

The relevance scores provided by trained semantic models can be used to search images based on their semantic assignment. Consider ranking images based on their relevance scores of semantic label ζ to be a single-semantic image retrieval, its performance can be evaluated using precision and recall [79] concepts.

$$precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \tag{5.5}$$

and

$$recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|} \tag{5.6}$$

In our case, they are defined similarly as

$$P = \frac{n}{N} \tag{5.7}$$

and

$$R = \frac{n}{|\iota_\zeta|} \tag{5.8}$$

, where n is the number of images labeled with semantic ζ in a list T_ζ of top N ranked images and $|\iota_\zeta| \leq |I|$ is the total number of images labeled with semantic ζ in database I. An image is considered relevant if it is labeled with query semantic ζ. Precision is the fraction of relevant images in result list T_ζ. Recall is the ratio of retrieved relevant images to the total number of relevant images in the database.

When an image database contains hundreds of thousands images, one wants to see a list of most relevant images instead of going through the entire collection. The more relevant images at the top positions in the list, the better the

retrieval. Precision-recall curve, which represents precision as a function of recall rate, can demonstrate how relevant images are distributed in a ranked list. Another evaluation measurement is mean average precision (MAP) score over 10 folds of experiment (eq. 5.9). The higher the MAP score, the more relevant images are retrieved at the top positions in the list.

$$AP_{\zeta} = \frac{1}{|\iota_{\zeta}|}\sum_{k=1}^{|I|} P(T_{\zeta}(I,k)) \tag{5.9}$$

Specifically, all database images are first ranked using relevance scores calculated by M_{ζ}. At each position k in the ranked list T_{ζ}, precision is calculated using *eq*. 5.7 where $N = k$ and n is the number of relevant images counted until cutoff k. When the k-th image is not relevant, P=o. The precisions at each position are then averaged to yield an AP score for semantic ζ. In this fold of modeling of M_{ζ}, an AP score is generated. Finally, the mean of AP scores for semantic ζ over 10 folds of modeling is calculated.

This semantic-based image search was then extended to include multiple semantic labels. The relevance scores for each semantic were used to calculate an overall relevance score as regard to a set Q of semantics selected by a user. Once the semantic models were trained, database images could be searched based on their relevance scores of multiple morphology semantics. For example, a set Q of query semantics is selected out of all N available semantics to query the database. Each image's overall relevance to this query is calculated using *eq*. 5.10. All database images are then ranked based on such relevance score. Top k most relevant images are eventually returned to the user (Figure 5.2).

$$s = \frac{s_{rel} * s_{irr}}{p * s_{rel} + (1 - p) * s_{irr}} \tag{5.10}$$

$$s_{rel} = \frac{1}{|Q|} \sum_{\substack{i=1 \\ i \in Q}}^{N} r_i \tag{5.11}$$

$$s_{irr} = 1 - \max_{i \notin Q} r_i \tag{5.12}$$

Specifically, the overall relevance score, s, is a weighted combination of average relevance score, s_{rel}, and the irrelevance score, s_{irr}. To calculate the average relevance score, s_{rel}, of an image, its relevance scores, r_i, for each of selected semantic in Q are averaged. The irrelevance score, s_{irr}, is the opposite of the maximal relevance score calculated for those semantics that are not in Q. In *eq*. 5.10, p is a system adjustment penalty to balance the scores of relevant and non-relevant semantics. It is heuristically set to be 0.002 in this study.

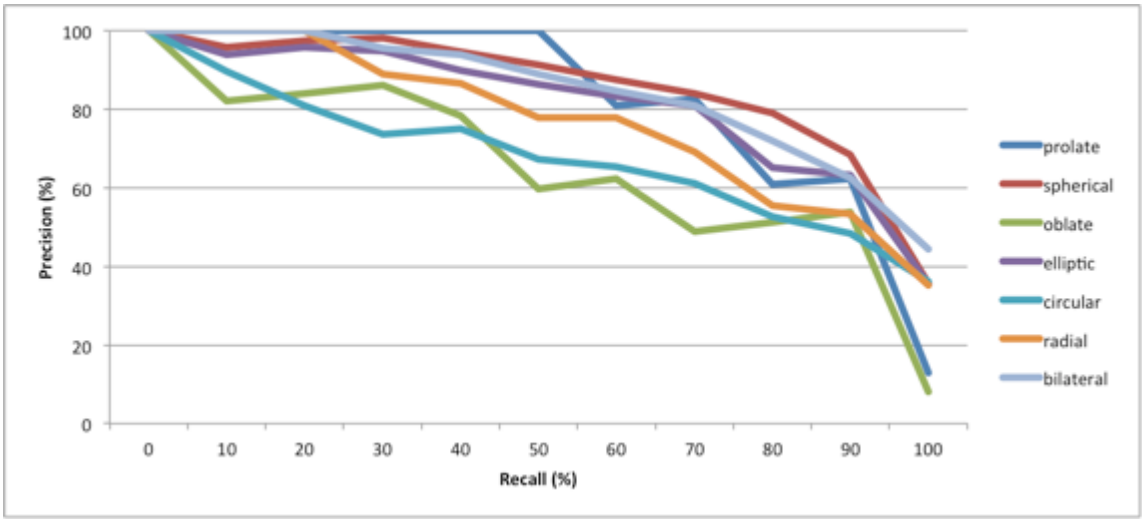


Figure 5.4 Precision-recall curves of morphology semantics for pollen images. As number of images retrieved increase, recall values gradually approach 100% while precision values gradually decreases since some non-relevant images are being retrieved.

The performance of the semantic-based image retrieval was evaluated using precision-recall curves and MAP scores. Figure 5.4 and Figure 5.5 are

precision-recall curves for morphology semantics of pollen and spores, respectively. In Figure 5.5 average precisions calculated from 10 folds of experiments were plotted as functions of recall for all 7 semantic models from 3 categories for spores. Precisions of all semantics maintained above 80% at 60% recall rate. For pollen images (Figure 5.4), even though precision-recall curves drop steeper, all precisions still maintained above 60% until recall rate of 60%. It is understandable that since pollen image samples in this study are distributed in 15 distinct taxa, it is more challenging for semantic models to find association rules of feature sub spaces that are generalized for image from all available species.

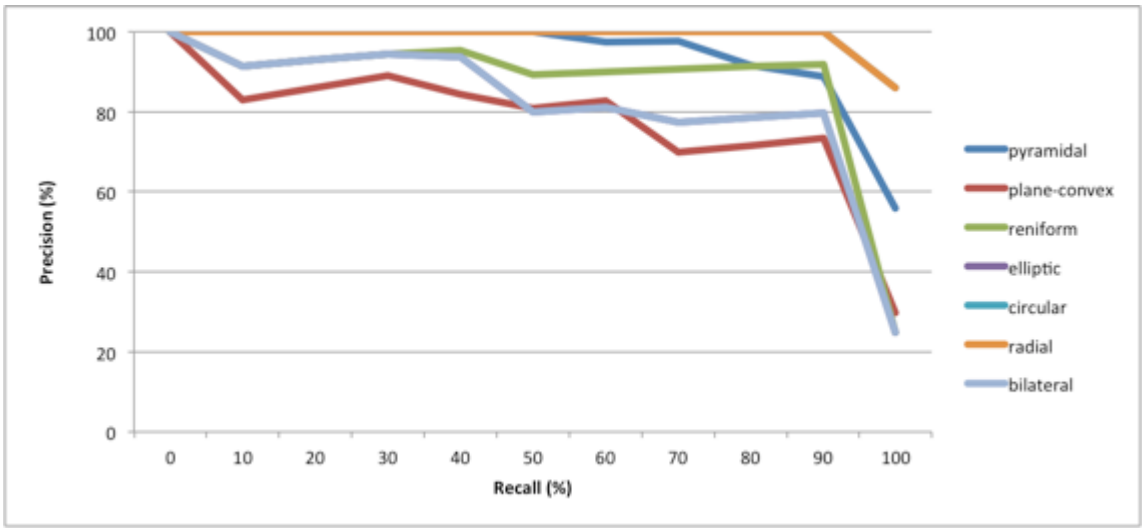


Figure 5.5 Precision-recall curves of morphology semantics for spore images.

Queries using pollen images had an average MAP score of 0.81/1.00 and queries using spore images had an average MAP score of 0.93/1.00 (Table 5.1). The average search time for pollen and spore images was less than 0.2 second and as short as 65 milliseconds.

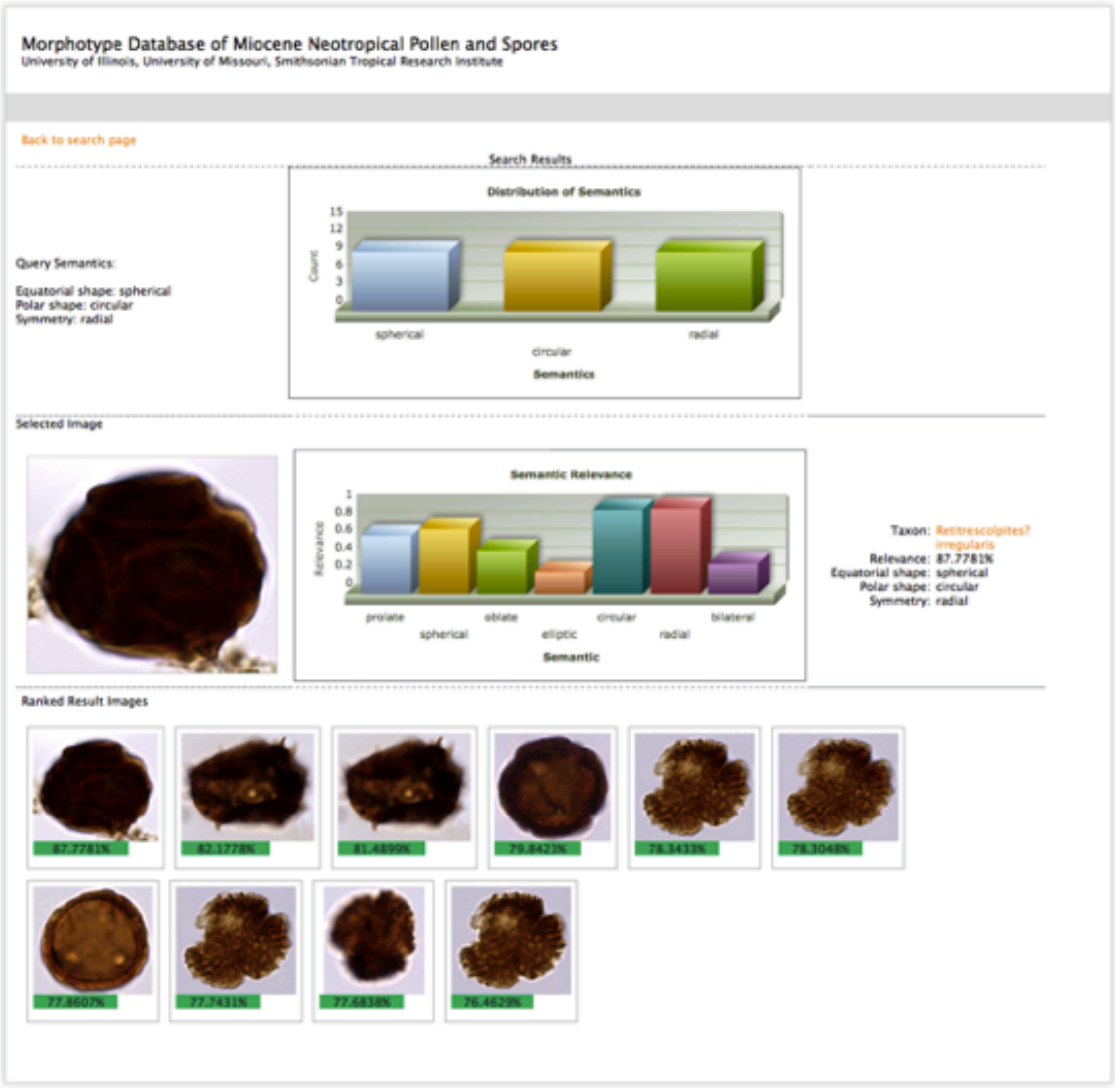


Figure 5.6 Searching for pollen images by morphology semantics. Top row: (left) Morphology semantics selected by user and (right) distribution of semantics in result images. Center row: (left) first image in ranked list, (middle) relevance scores calculated by trained semantic models and (right) additional information about this image, including taxon name, its overall relevance score as regard to user-selected semantics, its actual semantics annotated and stored in database. Comparing the actual semantics to relevance score chart, we can see that spherical has the higher relevance than prolate and oblate for equatorial shape semantic, circular is more relevance than elliptic for polar shape semantic, and radial is more relevant than bilateral for symmetry semantic. Bottom row: ranked result image list with their overall relevance score calculated using *eq*. 5.10 as regard to user-selected morphology semantics.

Figure 5.6 demonstrates the result page of searching pollen images using multiple morphology semantics. The semantic-based image search engine calculates overall relevance scores using *eq*. 5.10 based on each image's three relevance scores provided by semantic models for spherical equatorial view shape, circular polar view shape, and radial symmetry. The database images are ranked based on their calculated overall relevance scores and the top 10 most similar images are displayed. It is not required that retrieved images must have had all three morphology labeled. As long as their relevance scores are significant, the overall relevance still satisfied the search criteria.

5.3.3 Image Search using Image Examples

The semantic modeling additionally provides a basis for the query of images within the database and the retrieval of the most visually similar pollen grain images. In this way, a newly acquired image can be uploaded into the search engine to find similar types from the database. This allows for the comparison of morphotypes across analysts, potentially improving classification consistency among multiple experts.

The 276-dimensional features used in the initial morphological feature extraction formed a visual content space. These features were used to index the image database for fast retrievals. For simplicity, only three dimensions are depicted in Figure 5.7 to demonstrate the concept of content-based image retrieval from a multi-dimensional feature space.

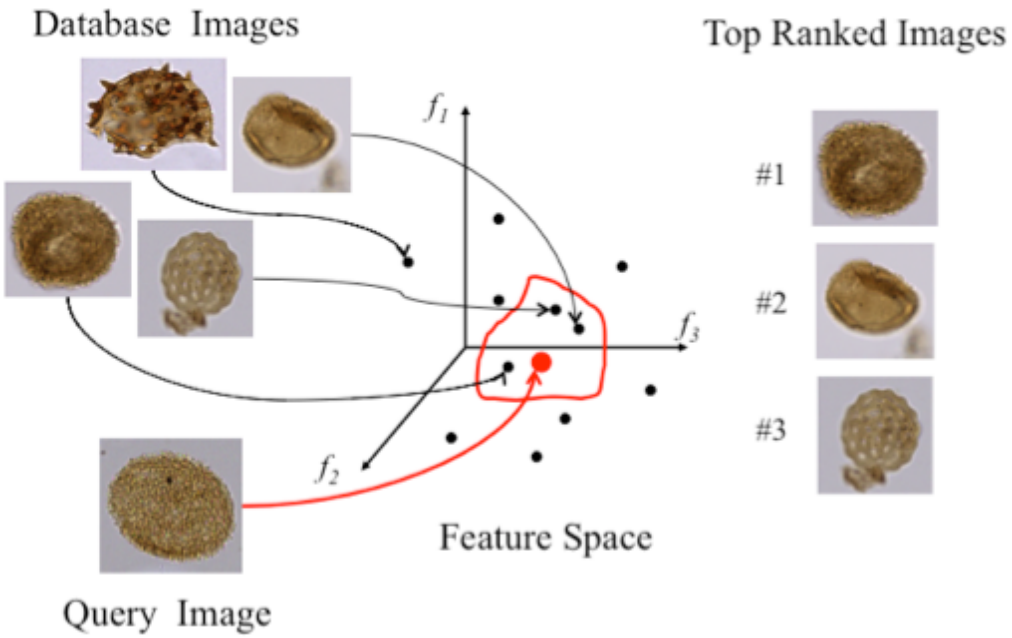


Figure 5.7 Hypothetical query by image example in a multiple-dimensional feature space (illustrated here in three dimensions). Each black dot represents a multi-dimensional feature vector that represents a database grain image. A query image is mapped into the same feature space (red dot). The nearest neighbors are selected and ranked with their corresponding images displayed to the user.

Each data point is a multi-dimensional vector representing an image in the database. The distance between a query point and a data point defines the similarity between the query image and a database image. In other words, the closer two data points are in the feature space, the more visually similar these two images are. With this defined similarity measure, images can then be ranked based on their similarity (distance) scores. For a large-scale image database (millions of grains), instead of exhaustively computing distances between the query image and all database images, customized database indexing structures, such as M-Tree [80] or EBS *kd*-tree [81], can drastically improve the efficiency of retrieval by strategically organizing the indexes of data in a high-dimensional

space. In our implementation, we created three M-Tree indexes by grouping the 276 visual features into colors, shapes, and textures.

Our system provides users with customized weighing options to search grain images. For example, one user may want to see images that are most similar to each other on shape features with less emphasis on color and texture variances. Color, in particular, is a highly variable characteristic as it is mainly controlled by the thermal maturation of the organic matter. Pollen grains are light yellow when thermal maturation is low but change to darker colors as rock maturation increases, reaching a fully black when organic matter is over-matured. In this scenario, users can customize their queries with a minimal weight on color index while emphasizing more on other two indexes.

In this study, content-based image retrieval was evaluated using precision in the top ranked images [79]. It is defined as the ratio of number of relevant images in top k ranked images (k = 10 in this study). A result image is relevant when it belongs to the same species as the query image example. Since the contribution of indexes to retrieval performance is not and should not be universally fixed across all species, we simulated possible combinations of weights, w, for color (w_c), shape (w_s), and texture (w_t) with an increment of 0.2 using labeled images in our database as training set.

In order to find the most suitable weight combinations, a series step was performed using permutated experiment results for each species in current dataset.

$$w \stackrel{\text{def}}{=} (w_c, w_s, w_t) \in \big((0,0,0), (1,1,1)\big] \tag{5.13}$$

1. Each image ι^{θ} from species θ was used as a query image for 215 $(= 6^3 - 1)$ times to search against the entire database. Precision p_i^{ι} $(1 \leq i \leq 215)$ were calculated for each query.
2. The weight combinations w_k^{ι} that produced the highest precision p_{max}^{ι} were identified for each query image, composing a set of candidates $W^{\iota} = \{w_k^{\iota} | p_k^{\iota} = p_{max}^{\iota}\}$.
3. For all images from the same species θ, their sets of weight combinations identified in step 2 were joined and the most frequently occurred combinations were considered candidates for most suitable weight choices. If there were multiple candidates with same number of occurrence, the one that yielded the highest average precision across all images in this species was considered the top choice.

The most suitable weight combinations were identified based on our current database image collection. Their values determined the retrieval precision of each query. Once the most suitable weight combinations were identified for each species, they were presented to users as the initial weights upon which emphases can be adjusted based on users' search preferences. As our database collection grows bigger to include more species and variety in morphology, new weight combinations can be learned to produce better retrieval results based on newly populated database.

On the content-based image retrieval interface (Figure 5.8), a user first picks an image as example and then adjusts emphasis on three trait semantic

categories using sliding bars. On the search result page (Figure 5.9), a list of top ranked database images that are most similar to image example is displayed.

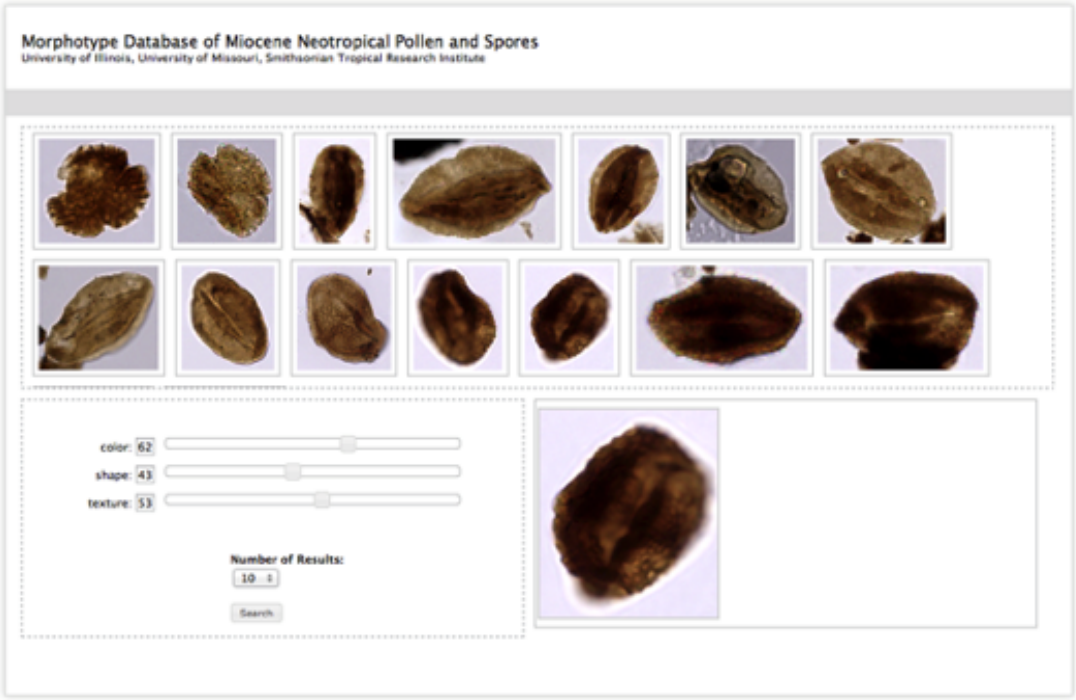


Figure 5.8 Searching for pollen images by query image example. Query example image is selected from example list and query weights on three indexes (color, shape, and texture) can be adjusted to user's preference. The weight values range from 0 (left end on the bar, representing no weight) to 100 (right end on the bar, representing the highest amount of emphasis).

CBIR is a much more complex image retrieval method than those by keywords and semantic labels. It is worth mentioning that species-level classification is the most challenging classification task in palynology [82] and images that are visually similar based on their content do not necessarily belong to the same taxon. As a result, even though a list of visually most similar images are ranked and returned, the precision value calculated by judging the variation of species can be much lower if multiple species are presented in the result list.

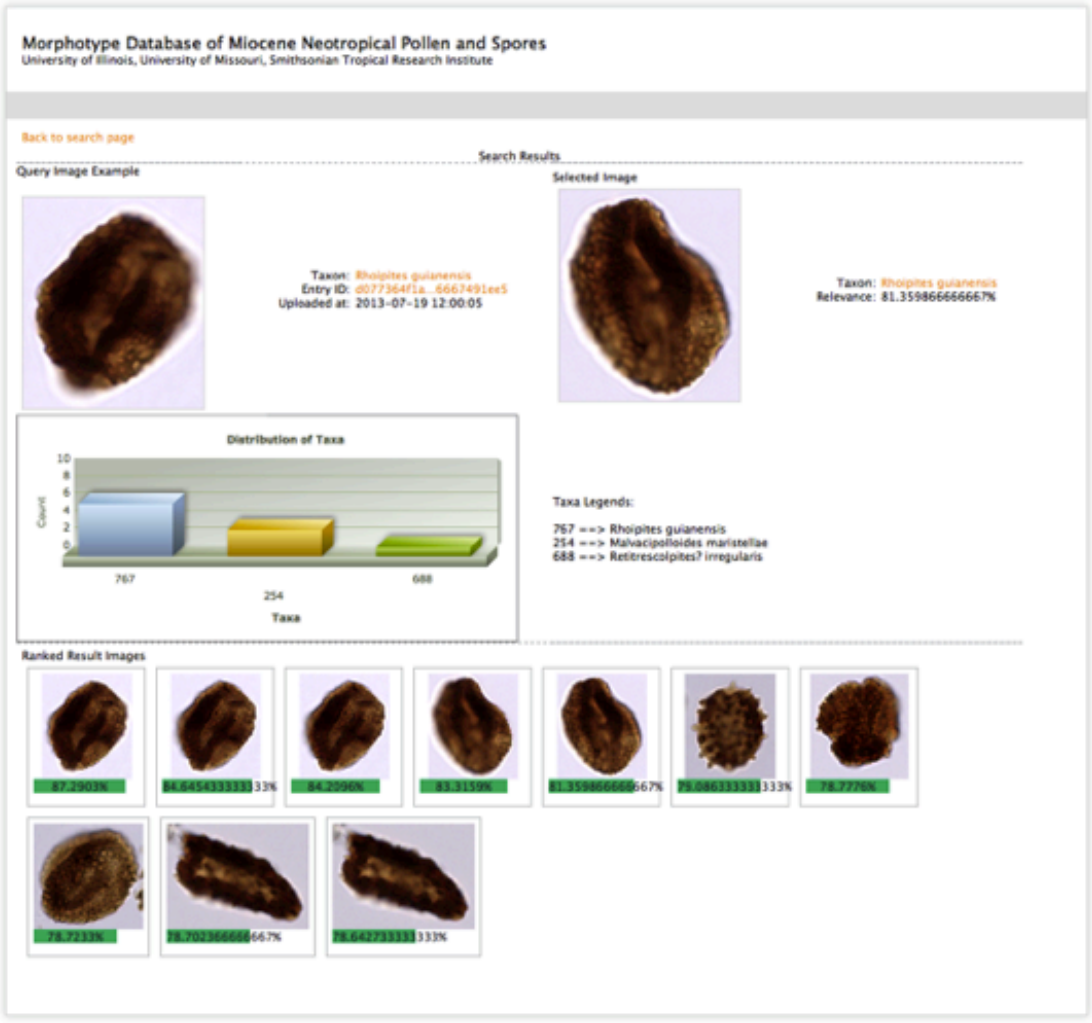


Figure 5.9 Search by pollen image example result page. Top row: (left) search example and (right) the fifth result in the ranked list. Center row: distribution of taxa count from results. Bottom row: ranked result image list with their similarity measures against the query image example (top-left). The bar chart in center row indicates that there is a mixture of taxa in the result images.

Table 5.1 lists the top 10 best-performing universal weight combinations that yielded average retrieval precisions of 57.8% and 72.3% for pollens and spores, respectively. This means that without treating species differently, using a one-fit-for-all weight combination is able to retrieve a list of pollen images in the database with 57.8% of them having the same species as the query image. The

retrieval precision using such universal weight combinations for spore images in the database is 72.3% on average.

Table 5.1 Top 10 best-performing weight combinations sharing similar retrieval performance for all pollen and spore species in the database.

Top 10 Performance with universal weight combinations	Pollen		Spore	
	Weights (color_shape_texture)	Average Precision	Weights (color_shape_texture)	Average Precision
1	0.4_0.6_0.6	57.9%	0.2_1.0_0.0	72.4%
2	0.8_1.0_0.8	57.9%	0.0_0.2_0.0	72.3%
3	0.6_1.0_0.8	57.9%	0.0_0.4_0.0	72.3%
4	0.6_0.6_0.8	57.8%	0.0_0.6_0.0	72.3%
5	0.2_0.2_0.2	57.8%	0.0_0.8_0.0	72.3%
6	0.4_0.4_0.4	57.8%	0.0_1.0_0.0	72.3%
7	0.6_0.6_0.6	57.8%	0.2_1.0_0.4	72.2%
8	0.8_0.8_0.8	57.8%	0.4_1.0_0.4	72.2%
9	1.0_1.0_1.0	57.8%	0.2_0.8_0.0	72.0%
10	0.2_0.4_0.4	57.8%	0.2_0.8_0.4	72.0%

Instead of choosing universal weight combinations for all images, choices were made for individual species. Table 5.2 shows the top choices of weight combinations for individual species and their average retrieval precisions. For images of some taxa, selected weights on trait semantics could be drastically different from others. For example, in order to get an average precision of 77.1% for all *Clavainaperturites microclavatus* (ID=1014) images, it is best to emphasize heavily on shape and reduce weights on color and texture. It is the most suitable weight combination choice to distinguish them from images of other taxa. For some species, one or two traits are ignored completely (weight is set to zero) to produce good retrieval results. This customized weight selections

become handy if the user has a small number of targeting species in mind during the search.

Table 5.2 Best-performing weight combinations for each taxon and their average retrieval precisions. Use taxon Retitricolpites simplex (ID = 722) as example. Every of its 24 images were used as query images and search against the database and retrieved back top 10 most similar images in feature space. All 215 weight combinations were used for each image yielding a total of 215*24 = 5160 queries. For each image as a query image, maximal precision was identified. There could be multiple weight combination (n/215) that produced same maximal precision for the same query image. All of these weight combinations were considered candidates. The candidates that occurred most frequently were final candidates. In this example, there were 4 (#Candidates) candidate weight combinations that were identified to produce maximal precisions in 9 (#Occurrence) query occasions, individually. The average precisions using each of 4 candidates across all 24 images were calculated and the candidate with the highest average precision was the top choice.

	ID	#Candidate	#Occurrence	Top choice	Ave. Precision
Pollen	722	4	9	0.8_0.2_0.0	50.8%
	767	2	17	0.4_0.6_0.2	73.1%
	688	7	6	0.0_0.4_0.2	63.4%
	570	2	24	0.0_0.6_0.2	59.1%
	571	21	10	1.0_0.6_0.2	74.7%
	552	1	16	0.2_0.8_1.0	73.1%
	451	8	14	0.2_0.0_0.0	61.2%
	511	1	9	0.2_0.0_1.0	68.9%
	450	1	20	1.0_0.4_0.8	67.2%
	254	5	16	0.2_0.0_0.0	49.6%
	1430	13	13	0.4_0.2_0.2	62.1%
	365	2	10	0.0_0.2_0.6	49.5%
	148	2	8	0.8_0.2_0.6	40.5%
	246	14	15	0.4_0.2_0.2	38.9%
	1014	3	11	0.2_1.0_0.2	77.1%
Spore	46	3	15	0.2_0.6_0.0	89.1%
	44	6	11	0.2_1.0_.00	65.0%
	282	2	12	0.2_0.2_0.4	78.4%
	45	9	8	0.2_0.4_0.2	58.3%
	43	5	24	0.0_0.2_0.0	91.1%

CHAPTER SIX

UTILIZATION OF BIG DATA TECHNOLOGIES IN PATHOLOGY INFORMATICS

In this chapter, we present our research in the field of pathology informatics, beginning at image processing and analysis, and ending with content-based image retrieval. We will also showcase the potentials of Big Data technologies that would improve the quality and capabilities of handling large-scale biomedical imagery.

6.1 Problems and Challenges

In the past decades, major advances in computer hardware as well as software technologies have helped numerous researchers and scientists to greatly improve their fields of expertise. This is also true in pathology. The examination of histopathological biopsy samples under microscopes is a crucial step in disease diagnosis. Conventionally, the characteristics of both cellular and gross phenotypic appearance of biopsied tissue samples are examined and summarized qualitatively by experts in their fields. This is also true for biologists who work closely with microscopic images of samples, such as the previously presented researches in mitochondria dynamics and pollen and spore grains from fossil samples.

During observation and diagnosis, pathologists examine an image slide from multiple levels of magnification, for example 2x, 8x, 20x, and 40x. On different levels of magnification, certain pathological patterns would be revealed on regions/objects of interest. On a coarse level, the dominant regions of interests are related to tissue-level structures, e.g. lymph node and its

components (capsules, primary and secondary follicles, sinus, etc.). On a finer level, follicular structures become more obvious, e.g. follicular center, mantle zones, marginal zones, inter-follicular regions, etc. On the finest level, individual cells and sub-cellular structures (chromatin pattern, nuclei, nucleoli, and various cytoplasmic changes) are the objects of interests. Their morphologies, textural patterns, as well as distribution and coverage are closely examined. The diagnosis is reached by examining the whole slide on multiple levels of magnification and a consensus of discoveries on these levels all contribute to the conclusion made by pathologists. The complex pathological content in digital slides and high-level reasoning process that pathologist utilize to reach diagnostic conclusions require years of professional experience and intensive training.

Such examination routines have potential limitations and pitfalls as this qualitatively visual examination may have inter- and intra-observer variability due to inconsistency of viewing environment, equipment adjustment, experience, fatigue induced from long hours, and extraneous external distractions, as well as the amount of data presented through microscopic examination. Poor reproducibility is not uncommon, and this may lead to inconsistency and difficulty in reaching conclusions for diagnosis and biological discoveries. The advance in human reasoning may be shadowed by the capability of memorizing huge amount of information observed and/or over-looked on the entire slide. This may potentially lead to unreliable diagnoses resulting in under- or over-treatment for patients and adverse consequences in quality of patient care and waste of money and resources. Therefore, a reliable, consistent, smart, and

efficient computer vision system may be appreciated for pathology studies as well as other fields that rely heavily on visual examination of images.

Thanks to the advances in computing hardware and software development in the field of computer vision, successful stories of computer-aided analysis have been presented in various biological and medical domains. However, as the imaging technologies advance, so does the volume and resolution of digitally scanned microscopic slides. An uncompressed virtual slide can reach several, if not dozens, of gigabytes in raw pixels. This puts a lot of pressure on the efficiency of image processing without losing the quality of the analysis results. Scientists and computer vision specialists are constantly searching for better ways to push the boundary of computational approaches, one of which is the Big Data analytics technologies.

The adaptation of Big Data framework into image analysis is, surprisingly, not as straightforward as one may envision based on the success stories in other research domains, such as health care [83, 84], business analytics [85], recommendation system [86], social networks [87], etc. Image analysis itself requires high-level domain knowledge that defines the underlying content of images and guides researchers to design computer-understandable programs to extract such content. The content of images captured is all stored hiding inside pixels. However, translating pixel data back from the other direction takes extra effort to achieve. For example, a market crowd is captured in a digital picture, and a computer vision program is designed to recognize human faces and flag salient subjects. The intelligence of facial recognition is only one of many examples that image analysis techniques can contribute to real-world problems.

As for medical image analysis, there is another realm of research topics that focus on tackling medical problems with computational strengths. The complexity and depth of image content is even higher in medical and biological fields. Rather than recognizing and comparing human faces, researchers are looking to discover biological objects, anatomical structures that bear pathologically meaningful visual patterns that eventually lead to diagnosis.

6.2 System Overview

There are multiple components in the complete pipeline of whole-slide image analysis and retrieval with Big Data infrastructure. In this section, we will briefly introduce the overall structure and its components.

Tile Extraction and Filtering: The original slide is first divided into individual non-overlapping tiles and saved for later analysis. This helps to reduce the total size of files to be examined, discarding non-meaningful regions on the glass slide.

Stain Un-mixing and Slide Normalization: To re-balance the color components across the entire slide collection, we utilize stain un-mixing methods introduced by [88] and [89] to first separate Hematoxylin- and Eosin-stains and then recombine them to result in a normalized color profile. This step not only minimizes the color bias from individual slides, it also provides us the single E-stain color channel that bears the major information this analysis is focusing on.

Multi-Scale Cell Identification: Image segmentation is one of the most difficult steps in the whole analysis process. In this particular study scenario, cell segmentation and identification is not an exception. We utilize a multi-scale

object segmentation method called differential morphological profile, DMP, with a customized post-segmentation consolidation to successfully identify cells with moderate overlapping and blurriness.

Rule-Based Cell Filtering and Refinement: Once preliminary segmentation is accomplished, we implement an extra step in an effort to maximize the quality of cell identification and the preservation of cell morphology. A set of domain-specific rules are constructed to filter unfit cell object candidates and to correct over- and under-segmented cell candidates at the same time.

Feature Extraction for Cells: A set of visual features are extracted for individual cell objects in this step to prepare for future reasoning based on cellular level information of which cell morphology is specially treated and extracted.

Profile Construction for Tiles: The content on individual tiles is also extracted by constructing a visual content profile using not only the low-level pixel information but also the special arrangement information of cells inside each tile. Treated as a non-directed graph, a group of properties are calculated to represent the organization pattern from the entire tile. The tile visual content profiles are used to perform high-level reasoning and knowledge discovery in the following steps.

Visual Category Discovery: Visual content is summarized into several categories based purely on the tile content represented by their profiles. The resolution of such categories can be adjusted as a parameter in tile clustering algorithms based on pathologists' preference.

In the next section, we will introduce each aforementioned component in our pipeline in details along with their preliminary results with explanations with domain-specific knowledge.

6.3 Component Details and Results

6.3.1 Tile Extraction and Filtering

The compressed whole-slide images in our collection are generated using Aperio® ImageScope® in TIFF format. Once uncompressed, the raw image pixels can reach several billions, some of which do not bear useful information. For example, a sliced and chemically treated thin tissue layer only convers a portion of the glass slide leaving the rest to be almost transparent which result in pixels that are close to white.

Additionally, due to the depth of image content and the complexity of subsequent image analysis methods, we choose to deploy a "divide-and-conquer" strategy to handle these large whole-slide images. This can be easily extended to other similar image modalities, for example, remote-sensing satellite imagery and space star atlas.

We choose to divide whole slide images into tiles of size 1000 by 1000 pixels based on the observation that when examining on the finest level of magnification, the viewing window, for example using Aperio® ImageScope®, on a computer monitor is roughly a rectangle with 1000+ pixels in both height and width. We mimic this practice and simplify the tiles as squares with 1000 pixels in each dimension. The Aperio® ScanScope® produces compressed digital whole slide images using TIFF standard. Our tile extraction program utilizes OpenSlide

library to extract individual tiles from such TIFF formatted slides without opening and decompressing the entire image. The following figures give us a rough concept on the scale of file size and pixel amount in WSI collections, using The Cancer Genome Atlas (TCGA) data set in our study.

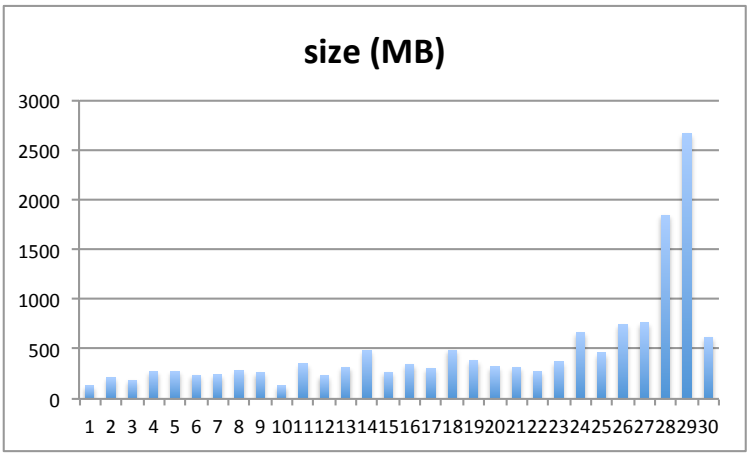


Figure 6.1 Tile sizes for each WSI images in TCGA data set.

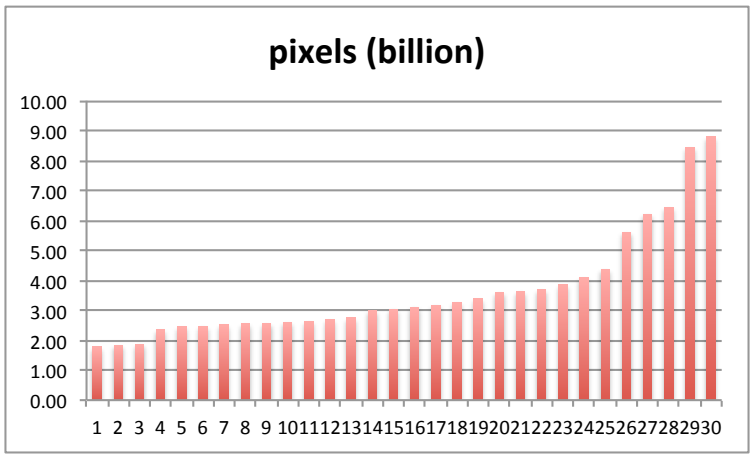


Figure 6.2 Total pixel count for each WSI images in TCGA data set.

6.3.1.1 Tile Extraction Efficiency and Scalability

Since we deploy a "divide-and-conquer" strategy to extract individual tiles, it is straightforward to parallelize such a process in a computer cluster. Using Apache Spark as the backbone and its **pipe()** function, we are able to initiate multiple instances of tile extraction program, written in C++, OpenCV, and OpenSlide, as an external program across multiple compute nodes with multiple CPUs (cores). To minimize the data traffic from worker nodes to driver node, we pass only calculated tile information to each workers where raw tiles are extracted and saved to hard disk directly. This method not only makes sense in the current step, it is mostly useful when we execute subsequent steps with much more complex analyses. The efficiency of this step is studied with different cluster configurations for a customized Spark program processing the TCGA data set. The tile size is predefined by a list of values, ranging from 1000 to 10000 pixels in both height and width. The program runtime is recorded using wall-time in seconds and illustrated in Figure 6.3. Although a more precise evaluation is to use actual program runtime instead of wall-time, it is not straightforward to obtain such runtime in a cluster environment with shared resources and complex scheduling mechanism. We argue that wall-time provides a close estimation of the actual execution time to explore the scalability scope.

A median-sized whole slide image can be divided into more than 2000 tiles. Using 10 compute nodes with 18 cores each, we can setup a Spark cluster with 1 driver node and 9 worker nodes. Consequently, we are able distribute 2000+ tile extraction instances across 162 cores (9*18=162). With such a configuration, a queue of tiles are extracted sequentially on each core and at the

peak performance, 162 tiles are being extracted simultaneously. As one can predict, the degree of parallelism would influence the overall performance of a Spark distributed computing job. The contributing factors include, but not limited to, the following list:

- Number of worker nodes
- Number of CPUs/cores per worker node
- Memory allocated on each node
- CPU frequency, e.g. 1.2 GHz in our current cluster
- External program runtime, especially the longest runtime on worker cores
- Distribution of tasks across the cores

In general, as the pixel count increases, so does the runtime. This trend is also observed as tile size increases. For some individual executions, the runtime increases more quickly. When the tile size exceeds 5000*5000, the runtime increases more drastically in general. This is due to longer runtime for individual external tile extraction program. However, we also observe a few decreasing cases for bigger tile sizes. This is due to two competing factors: external program runtime and number of active cores. Specifically, when tile size increases the total number of tiles decreases. In such situation, although it may take longer to run each tile extraction, the task queues distributed across cores may become shorter and some cores may be idle with no task queue assigned to them. In addition to Spark's own task scheduling mechanism that tries its best to balance the distribution of tasks, we also utilize data repartitioning to mitigate such effect.

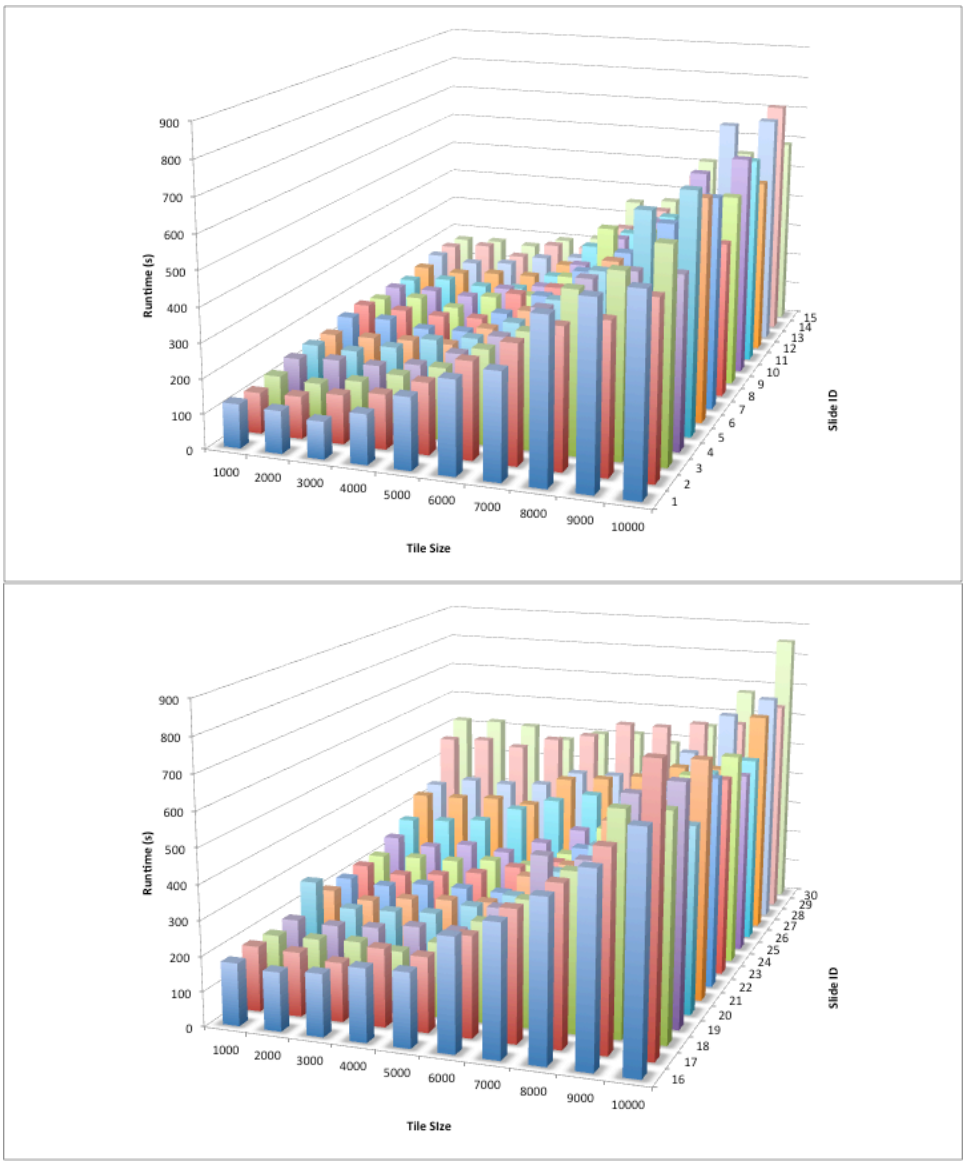


Figure 6.3 Scalability tests for tile extraction on TCGA data set. (top: slides #1 to #15; bottom: slides #16 to #30)

The scalability can also be shown in the following quartile box chart as compared to the pixel count tread shown in Figure 6.2. As the pixel count increases close to a 5 times difference (1.8 billion to 8.8 billion), the median runtime only increases 2.2 times (235.9 s to 514.9 s).

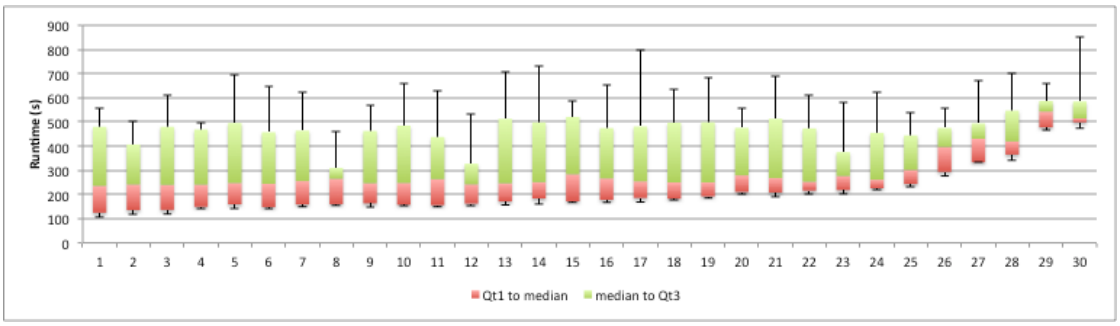


Figure 6.4 Tile execution runtime statistics for individual slides.

6.3.2 Stain Un-mixing and Slide Color Normalization

The raw whole slide images are captured in the RGB color space as for most of common image capturing devices. In image processing, several color systems can be used to represent different color components (Figure 6.5), for example RGB, CIE L*a*b, and HSV. However, the RGB color system is not the most suitable color space that reveals the true visual patterns residing in the imagery. After preliminary experiments, we chose to use the methods presented in [88] and [89] to separate stains in H&E virtual slides. The assumption on which this strategy is based is that the Hematoxylin (blue) and Eosin (pink) stains can be independently captured and thus are separable The details of this method are beyond the scope of this project. We will briefly explain the basic concepts here. For in-depth discussion, please refer to [88]. The color-unmixing method was implemented in Matlab and provided by authors of [90].

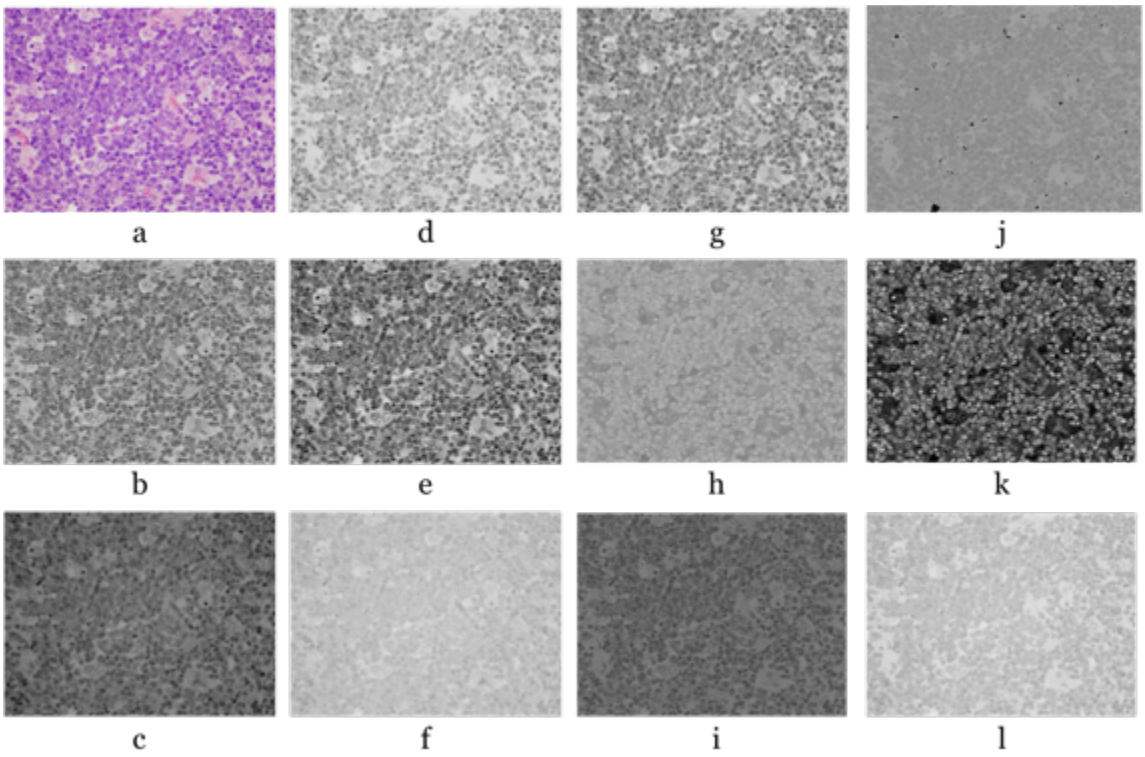


Figure 6.5 Demonstrate different color spaces. Original image (a), H-stain (b), and E-stain (c); R (d), G (e), and B (f) channels; L (g), *a (h), and *b (i) channels; and c: H (j), S (k), and V (l) channels.

Similar to tile extraction, stain un-mixing can also be performed in parallel. In practice, we use a Job Array provided by SLURM (Simple Linux Utility for Resource Management) [91] to deploy multiple Matlab instances independently. It is important to point out that this stain un-mixing does not always produce an actual separation due to constrains in mathematical calculation, such as singular value decomposition (SVD) and matrix operations. The most common reason is that the raw tile images are from the regions that are not stained (transparent in light, appearing in pseudo-white in digital images). In this situation, the model assumption that pixels are comprised with H-stain and E-stain related color components is not satisfied. This in fact helps us to trim down the overall number of tiles to be analyzed in the subsequent steps.

The subsequent analysis steps use the separated H-stain image to identify cells and extract tile content profiles. The reason behind this choice as explained in [89] is that 1) Hematoxylin stain primarily highlights the chromatin and chromosomes and it in turn brings out the morphology of nuclei and 2) pathologists closely examine the morphology of nuclei, predominantly, with some exceptions such as for nuclear/cytoplasmic ratio analysis.

Another merit of stain un-mixing is that the recombined H- and E-stains result in a normalized image. In common practice, pathologists examine digital slides one at a time. In other words, the discrepancies between slides are not crucial for patient-based diagnosis. With professional training and clinical practice, such discrepancies do not influence the overall observation and the reasoning to backup final diagnosis. However, such a scenario changes once the computational component has joined the diagnostic process. High-throughput computation technology provides the ability to batch-process a collection of digital slides. At this point, the visual difference would introduce bias in image analysis. Therefore, it is crucial to normalize the appearance across the entire collection of slides, eliminating the skewed values from one slide to another. Figure 6.6 gives an example on how a raw H&E image tile can be separated into Hematoxylin- and Eosin-stain images and then recombined to a normalized result.

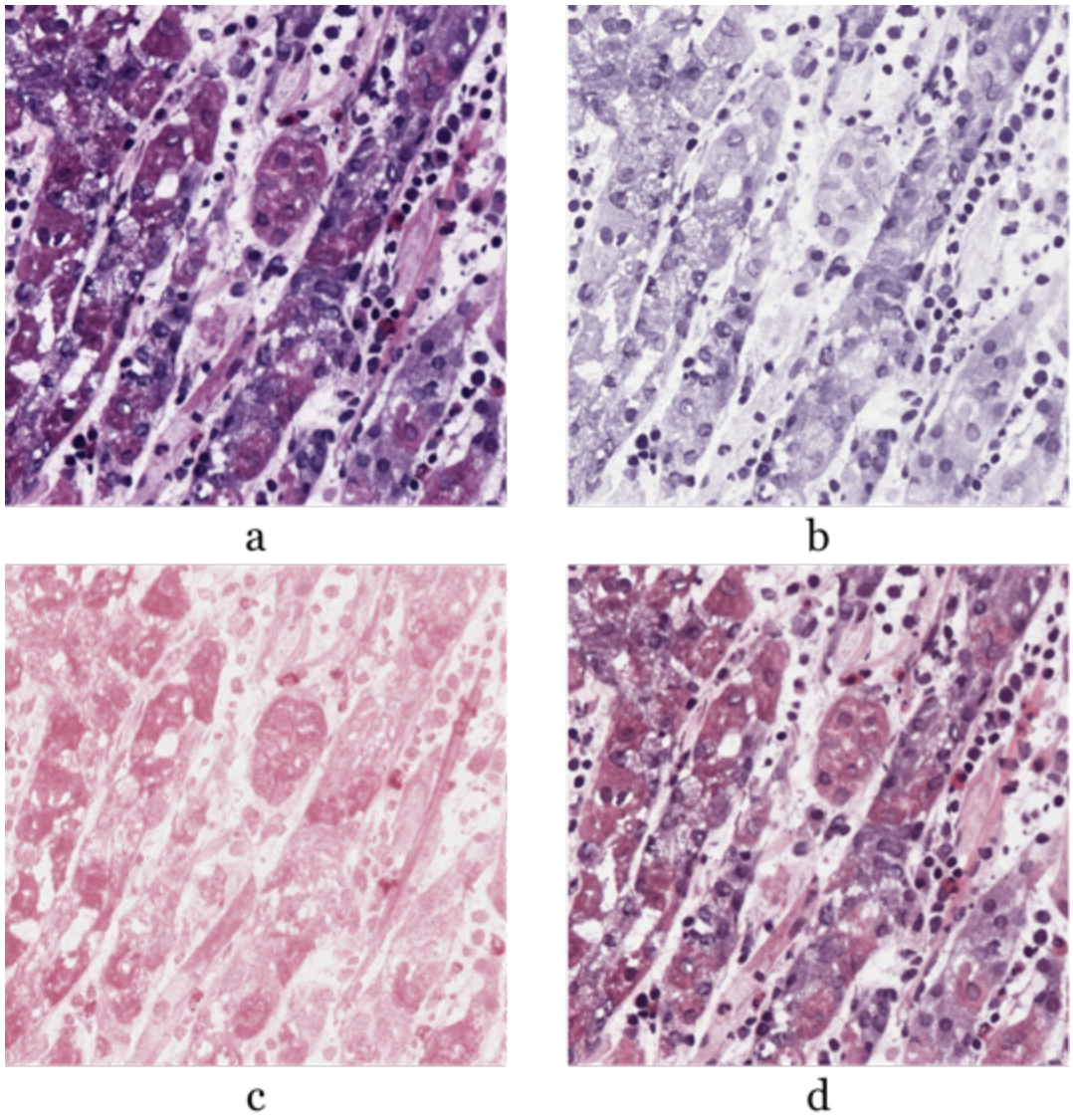


Figure 6.6 stain un-mixing. The (a) raw image was first split into two stain color channels: (b) H-stain and (c) E-stain and then recombined to obtain a normalized image (d).

In another angle of explanation, one would prematurely reason that we are leaving out the whole color spectrum in the slides, such as luminance and color saturation. As it turns out, the saturation does not reflect the "degree of reaction" between color-labeled antibodies and their targeted antigens in cells. The degree of saturation of stained cellular structures does not linearly indicate the amount

of interaction between color-labeled antibodies and their targeted antigens to highlight the cellular structures.

In this study, we are focusing on identification of cells (mostly cell nuclei) for positive detection of abnormalities reflecting the presence of certain diseases. Therefore, our priority drives us to choose H-stain for further analysis. Yet, we do not eliminate the contribution of E-stain as well as other components from different color spaces.

6.3.3 Cell Identification using cDMP

As explained in the background section, differential morphology profiles (DMP) has its advantage on smartly revealing objects of interest and reflecting both their morphology (size and shape mainly) and intensity on the spectrum (single channel mainly). We implemented the hybrid version of DMP adapted from [11] and previous works in [92]. After one initial morphological operation to build mask image, and two raster scans, one forward and on backward order, to perform morphological reconstruction, a result image is obtained using a fixed-size structuring element (SE) that is defined in both size and shape. In order to capture signatures from multiple scales, a list of SE's are selected and therefore produce a stack of reconstructed images.

Following the morphological reconstruction, the pairwise differences on pixels values between any two adjacent scales – the derivative or difference (the D in DMP) – are calculated. This difference exposes the peak response of reconstruction, which is considered the characteristics of object that this particular pixel resides in. The farther this peak is on the spectrum indicates that

the object this pixel belongs to has a relatively larger size. In addition, the side on which this peak is at on the profile scale also indicates the intensity of the object this pixel belongs to. One side is for “darker” objects (resulted from closing by morphological reconstruction) and the other is for “brighter” objects (resulted from opening by morphological reconstruction).

There are different strategies to consolidate the stack of reconstructed images in the literature [92, 93, 94, 95]. Most of the time, the explanations are relatively brief. In general, they either examine overlapping objects identified from different scales and eliminate reoccurrence [92], or treat the objects separately and use customized filtering to make selection [89]. In our empirical observations, the effect from different SE sizes introduces incremental results in the reconstructed image rather than a clear differentiation on the object boundaries. We lean toward keeping most influential information from each scale and consolidate them into a single image. That being the characteristic (peak response from DMP operation) of each pixel – we call it the consolidated DMP image (or cDMP image).

Once the cDMP image is constructed, an image binarization needs to be completed in order to reveal the boundaries of objects. The first choice to binarize a grayscale image is to use Otsu thresholding. However, the boundaries of cell nuclei in a cDMP image are in fact gradually easing into the background instead of a clear-cut differentiation. Additionally, a simple experiment reveals that the pixel value distribution does not fit the strong assumption that Otsu uses – pixels distributed generally into two peaks in grayscale histogram. Instead of Otsu [8], we use the Adaptive Thresholding [96] to differentiate foreground (cell nuclei)

from background (bright light) while preserving the morphology as much as possible. The comparison of two thresholding methods on a H-stain image with a prominent single-peak histogram is presented in Figure 6.7.

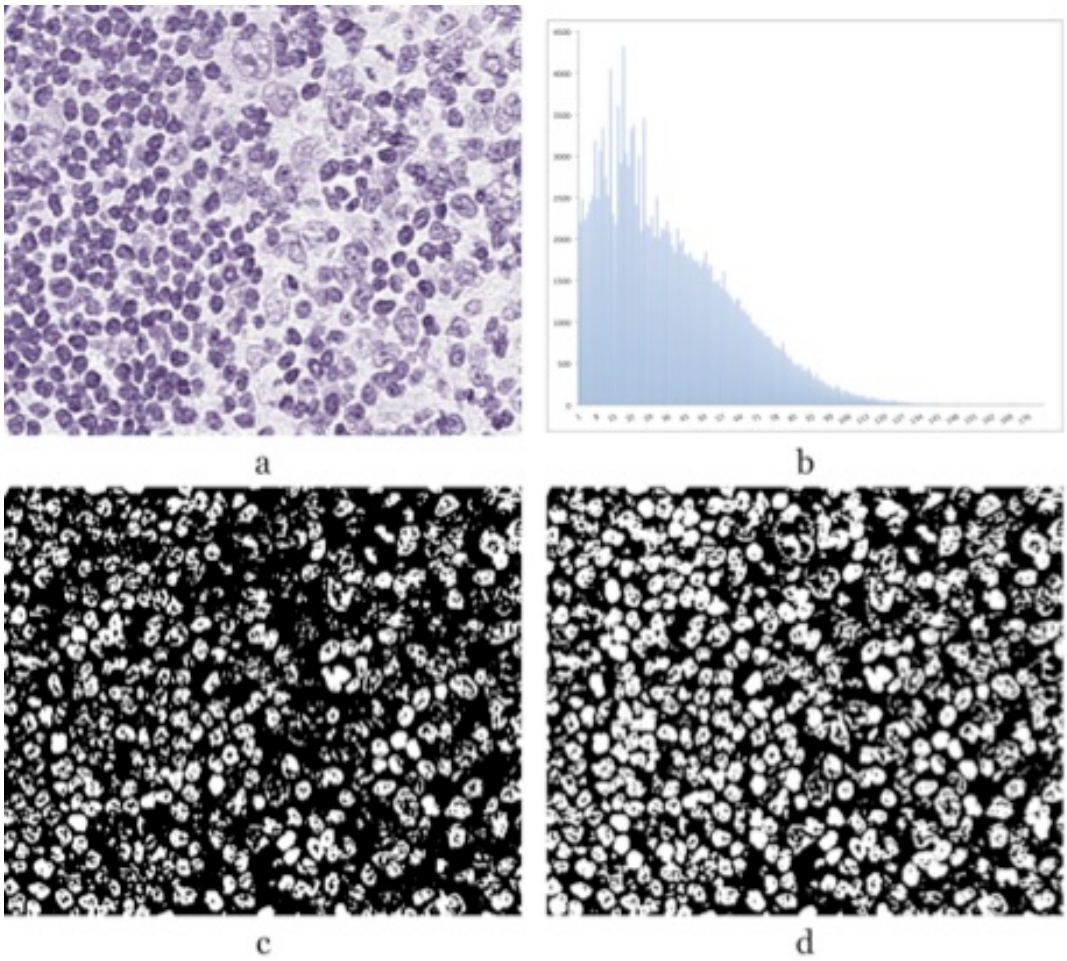


Figure 6.7 Comparison of two thresholding methods. (a) H-stain image, (b) histogram of (a), (c) result image by Otsu thresholding, and (d) result image by Adaptive Thresholding. Note: grayscale histogram only show pixel value bins from 1 to 182. Bin 0 value is 132711 and bins from 183 to 255 are all zeros.

After the binarization of cDMP image, some cell nuclei can still be touching/overlapping. We address this issue with marker-guided watersheding. There are numerous applications that discuss the utilization and variation of watersheding method. Despite the variations, researchers agree that roughly

defined markers would greatly improve the segmentation results. In this study, we define markers using lightly thresholded geodesic transform.

6.3.4 Rule-Based Cell Filtering and Refinement

Sometimes, Watersheding still couldn't segment closely clustered irregular objects. We then convey high-level concepts into median-level morphology-based rules to iteratively break down cell clusters. The set of rules can be adjusted for different use cases. In our study, cell nuclei are relatively round, from circle to ellipse with some exceptions that show "twisted" or "elongated" contour.

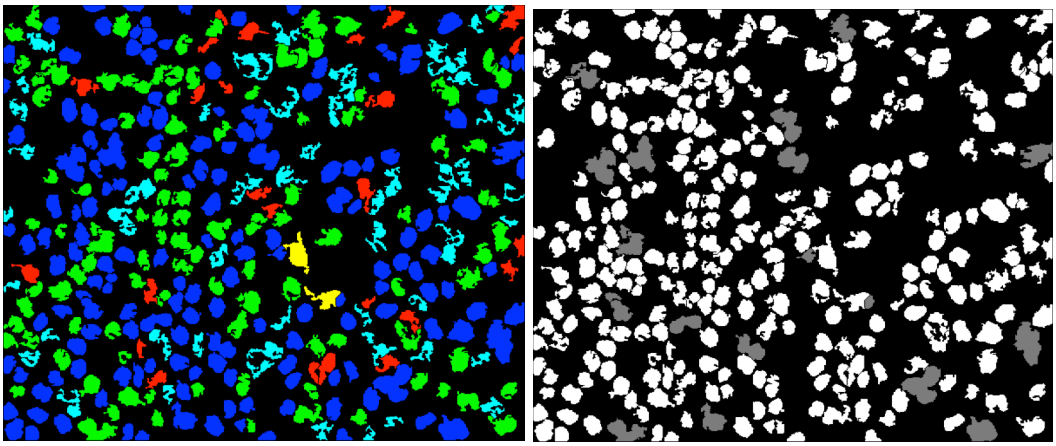


Figure 6.8 Iterative cell filtering and refinement. (Left: identified cell candidates, blue: positive, green: minor distortion, red: uncertain, cyan: discard, and yellow: clustered; Right: final segmentation results, white: positive, and gray: minor uncertainty).

We used the following rules to make decision on whether a candidate is 1) truly a nucleus with higher confidence, or 2) a cluster of a few nuclei that are separable with further watersheding, or 3) irregular objects that are hardly possible nuclei. This process is performed in iteration until 1) all objects are determined to be either nuclei or artifact and/or 2) iterations reach an upper limit – when further watersheding is of minimal influence in overall

segmentation results. Figure 6.8 shows results of iterative cell filtering and refinement.

6.3.5 Cell Feature Extraction

To describe the morphological patterns of cell nuclei, we need to extract numerical features from identified nuclei. There are a wide variety of features choices. The selection of features varies from application to application. As stated in a previous section, the stain concentration (a.k.a darkness) does not linearly reflect the degree of staining. Therefore, we treat the isolated H-stain images as grey scale images of interest. Moreover, morphological characteristics are used to describe cell nuclei instead of color information, especially in our study. On the other hand, this does not hinder us from developing a useful pipeline of algorithms from identifying cell nuclei to visual feature extraction, from tile clustering to slide retrieval. With domain experts' insights, we can also weight differently the subset of features based on their contribution to diagnosis.

With this point addressed, we are looking at what features would best describe the morphological characteristics of cell nuclei that we have discovered in previous steps. The pixel intensity features are chosen to be: Otsu threshold, mean and standard deviation, and binned histogram. The gray level co-occurrence matrix (GLCM) is used to calculate a set of textural features averaged over different direction [4]. Shape-related features include: Hu moments, aspect ratio, compactness, convexity, form factor, roundness, solidity, perimeter, and mass center.

6.3.6 Tile Content Profile Construction

Similar to the feature extraction step for cell nuclei, we also select a collection of features that best describe the visual patterns in image tiles. First of all, the same set of features is calculated to describe the global tile content from pixel values, for example, mean and standard deviation, histogram, and GLCM textures. Secondly, instead of morphological features, as for cell nuclei, we calculate graph-based properties from the network of cells inside the tile.

The distribution of cells is another checkpoint that pathologists refer to when analyzing the virtual slides. In order to represent such distribution, we use the mass centers of cell nuclei, which were calculated in the previous step, to construct a network of cells. Two widely used graph structures are constructed, namely the Delaunay Triangulation and the Voronoi Diagrams. Figure 6.9 demonstrates the visual differences between two tiles; one is showing generally even distribution of cells while the other tile has cells that are lined up in sheets. As presented in the Delaunay triangulations, pairwise distances (triangle edges) are short in the first tile as compared to the second tile with bigger distances between cells from different sheets. Since Voronoi diagram is derived from Delaunay triangles, we also observe the differences in polygon sizes and distribution.

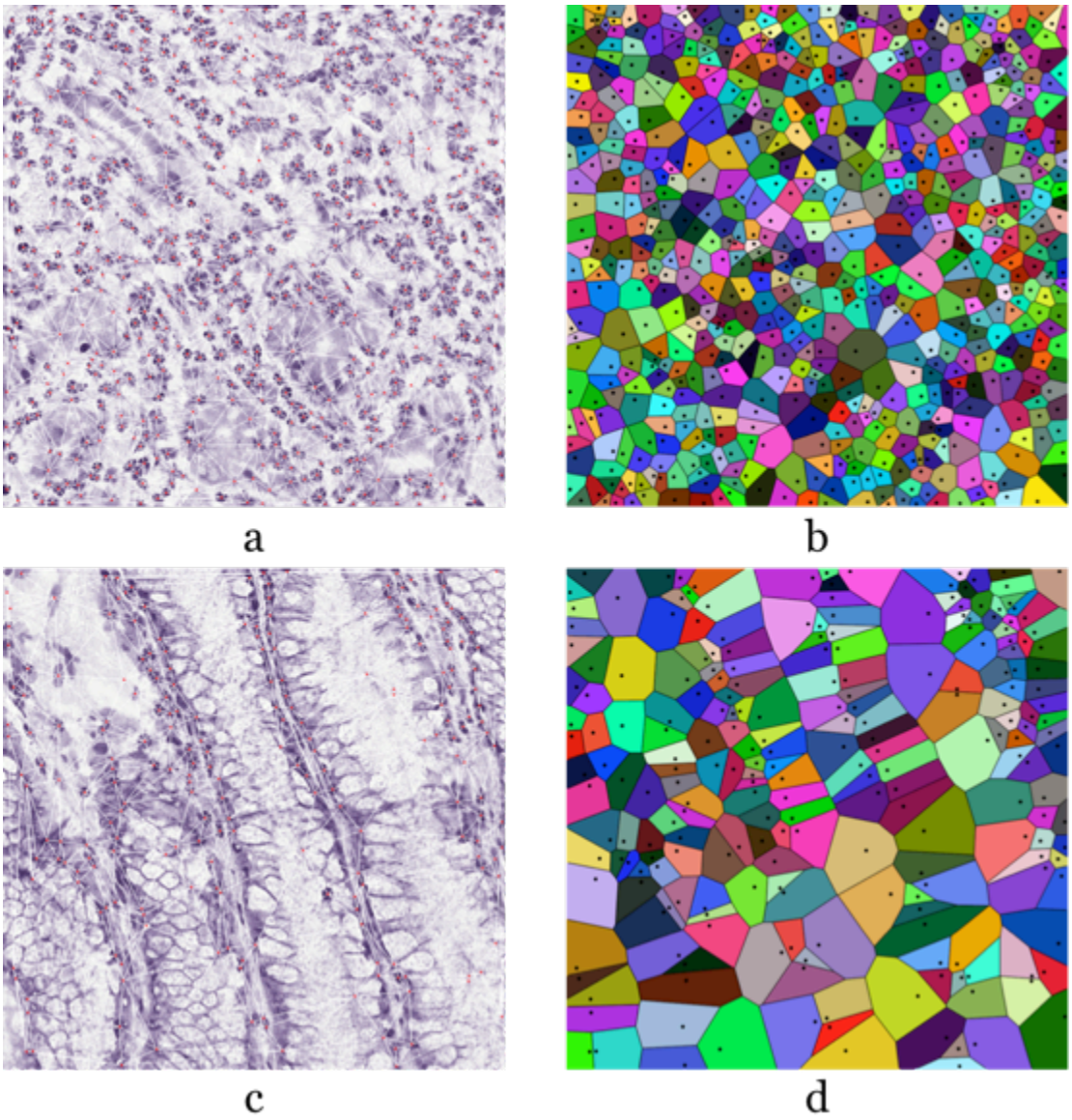


Figure 6.9 Examples of Delaunay and Voronoi structures of tiles with distinct patterns.

To quantitatively describe the distribution pattern using these two graphs, a set of properties are calculated (Table 6.1). The differences between the two example tile patterns are shown again in Figure 6.10 and Figure 6.11 with histogram representations of Delaunay triangle edge sizes and Voronoi polygon area sizes.

Table 6.1 Graph properties and their calculations for both graphs.

Delaunay Triangulation	**Voronoi Diagram**
Mean and standard deviation of pairwise distances between cell nuclei	Mean and standard deviation of cell-centered polygons
Ratio of min and max pairwise distance	Ratio of min and max polygon sizes
Binned histogram of pairwise distances	Binned histogram of polygon sizes

We can see that, in Figure 6.10 for evenly distributed cells in tile a, the triangle edge size distribution is close to a normal distribution with a slightly longer "tail" on the upper end. This reflects the homogeneous pairwise distances with slightly bigger gaps for those cells that are clustered as circles, leaving some spaces inside. On the other hand, sheets pattern results in a "two-peak" phenomenon showing a relatively large portion of cell pairs that are far from each other. Similar conclusions can be drawn from Voronoi polygons size distributions in Figure 6.11.

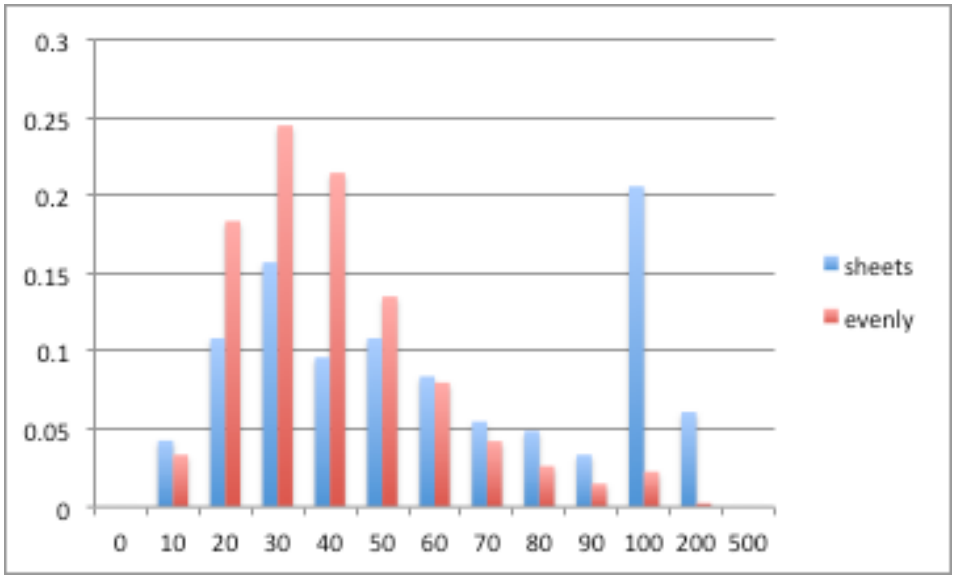


Figure 6.10 Edge length distribution from Delaunay triangulation graphs in two types of tile patterns.

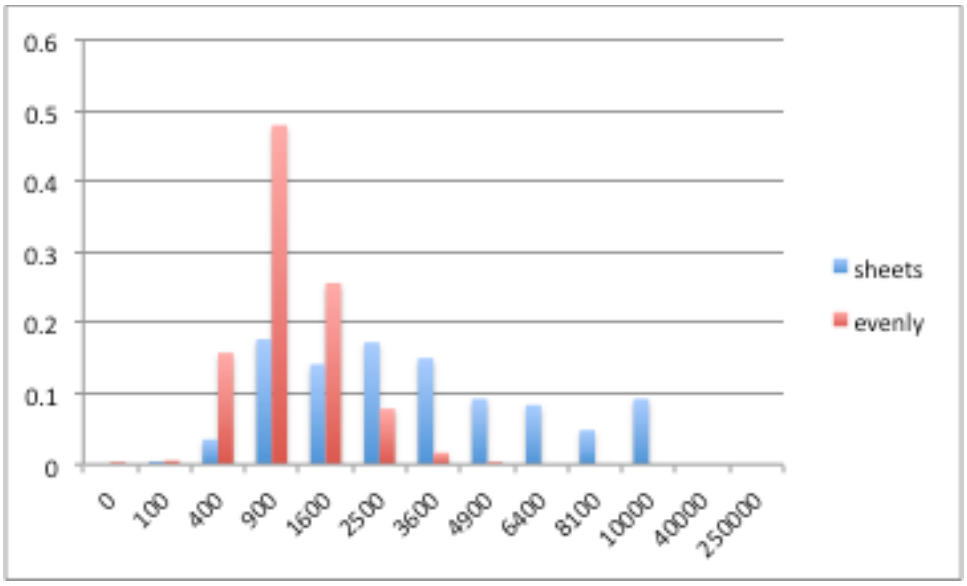


Figure 6.11 Polygon size (in pixels) from Voronoi diagram in two types of tile patterns.

Here again, we test the scalability of Spark tile profile construction. Specifically, six WSI examples are selected from TCGA data set: two smaller slides (#1 and #2), two median-sized slides (#15 and #16), and two larger slides (#29, #30). With total number of cores set to be 32, 64, and 128. We arranged different configurations with number of nodes in the cluster and number of cores per node.

Instead of tile extraction, we tested the configurations using tile profile construction for this set of experiments. The execution wall-time is used for all six configurations shown in Figure 6.12. In general, the more the total cores the shorter the runtime, and the smaller the file size the quicker the program finishes its execution. With a fixed number of total cores, Spark cluster favors a bigger number of cores per node due to less inter-node communications. On the other hand, this observation weakens when total core count reached 64. This indicates that the inter-node communication cost is counter-balanced by the cost of longer

task queues on each node. As the number of total cores multiplied 4 times from 32 to 128, the execution time shortens as big as 6 times (slide #15). Based on these observations, we can make the following suggestions on Spark cluster configurations for image processing.

- When resources are sufficient, allocate as many computing cores as possible.
- When resources are constrained, limit the number of nodes while maximizing number of cores per node may improve the runtime performance.

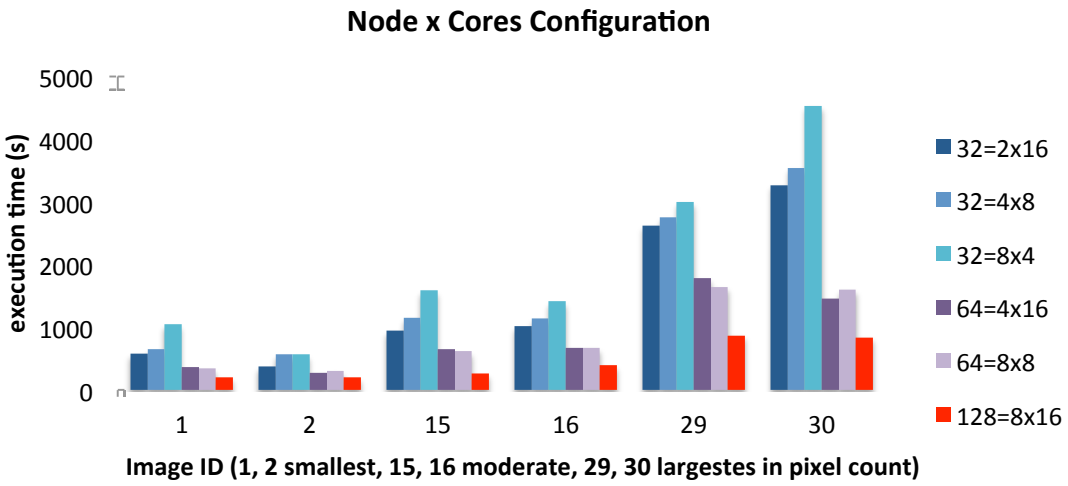


Figure 6.12 Scalability test on node-core configurations.

6.3.7 Visual Category Discovery

Clustering is one of the most widely used machine learning techniques. It groups data points into clusters based on different criteria trying to minimize intra-cluster difference while maximizing inter-cluster differences. The difference, sometimes called distance, is defined slightly differently from one

variation to another. The most commonly used distance calculations are, but not limited to, Euclidean distance, Hamming distance, Lp-norm, etc. Clustering is also considered as an unsupervised machine learning techniques. It does not require data points being labeled with predefined categories. However, in most of the variations, it requires a predefined number (*K*) of clusters with some exceptions, such as MeanShift [97]. When a research domain has a clearly targeted number of clusters, it is relatively easy to determine the *K* value. On the other hand, as suggested by [98], the choices of an optimal *K* value can sometimes be loosened to more than one candidate, representing the resolution of clusters in a hierarchical relationship.

Apache Spark has its own machine learning library, MLlib, which provides readily available packages of machine learning algorithms that are mostly published variations suitable for parallel computation. Our application in this step is looking for clusters that represent visual patterns, in groups, from tiles within and across individual whole-slide images. The value of *K* is difficult to determine. On the other hand, the number of different patterns is usually limited based on our observations. The choice of *K* can be determined by a combination of consulting domain experts and experiments. Keep in mind that, we are looking for clusters of closely similar tiles in the goal of finding a finite and limited number of representative tiles to describe the overall patterns inside a WSI. In other words, the major contribution of tile clustering is to drastically reduce the number of tiles to be closely examined by domain experts without losing crucial patterns that are rather not dominant nor salient and therefore easy to be overlooked by human observation.

In this study, we use the Apache Spark MLlib's K-means package that is developed based on the work [99] of K-means|| (" || " is pronounced as pipe). The value of *K* is initially determined to be 6, empirically.

Each tile is represented as a high-dimensional feature vector using features introduced in previous step. K-Means|| analyzes thousands of tile feature vectors and produces *K* cluster centroids with the same dimension as tile feature vectors. However, they may not necessarily be an actual data point. Therefore, we need to further determine for each data point where they belong to by comparing their distance to each of *K* centroids. The centroid that is the closest is the where this data point should be grouped into.

Table 6.2 Follicular Lymphoma data set grades.

Image ID	FL Grade	Image ID	FL Grade	Image ID	FL Grade
1	III-A	5	I	9	I-II
2	III-A	6	I	10	I-II
3	III-B	7	I	11	I-II
4	III-B	8	I		

Once all the tiles are assigned with a cluster label, the whole slide images are represented with tiles falling under *K* visual categories, proportionally. Figures 6.13 and 6.14 display the cluster distribution percentage within each slide with representative tiles from each cluster, *K*=6 and *K*=4 respectively. The dataset used in this experiment is a collection of 11 whole-slide images from patients with Follicular Lymphoma. The only information we have other than the images themselves are the final diagnosis with a WHO grade. Case diagnoses are summarized in Table 6.2. Three slides (#3, #4, and #5) were subsequently

removed from this dataset according to pathologist's suggestion. From Figure 6.13, we have drawn the following observational points.

- Clusters 2 and 6 show similar pattern that is described with term "starry sky" by some pathologists for diagnosis. However, the pattern is more obvious in cluster 2 than cluster 6, in which normal lymphocytes are the majority leaving fewer hollow areas.
- Grade I slides (#6, #7, and #8) show almost exclusive occurrence of tiles from cluster 4, which demonstrate a mixture of fibrous tissue and small centrocytes.
- Grade I-II slides are represented with distinguishable cluster percentage pattern – dominating number of tiles from cluster 6.
- Slide #9 also has a great portion of tiles from cluster 2. This gives us the indication that we should take a closer look at all three slides to see whether slides #9 should be assigned with Grade II since it has a large portion of tiles (cluster 2) with more advanced "starry sky" pattern as compared to cluster 6.
- Slide #1, although assigned with Grade IIIA, has a similarity in tile cluster distribution to slides #9 and #10. This intrigues us to question whether it is a case being over-diagnosed. In other words, slides #1 and #9 should be further examined side-by-side to see what common pattern they share and what set them apart.
- Cluster 3 shows a much higher percentage in slide #6 (Grade I) as compared slides #7 and #8. This observation would warn us that there

might be an uncertainty on the diagnosis of slide #6 that needs a second opinion.

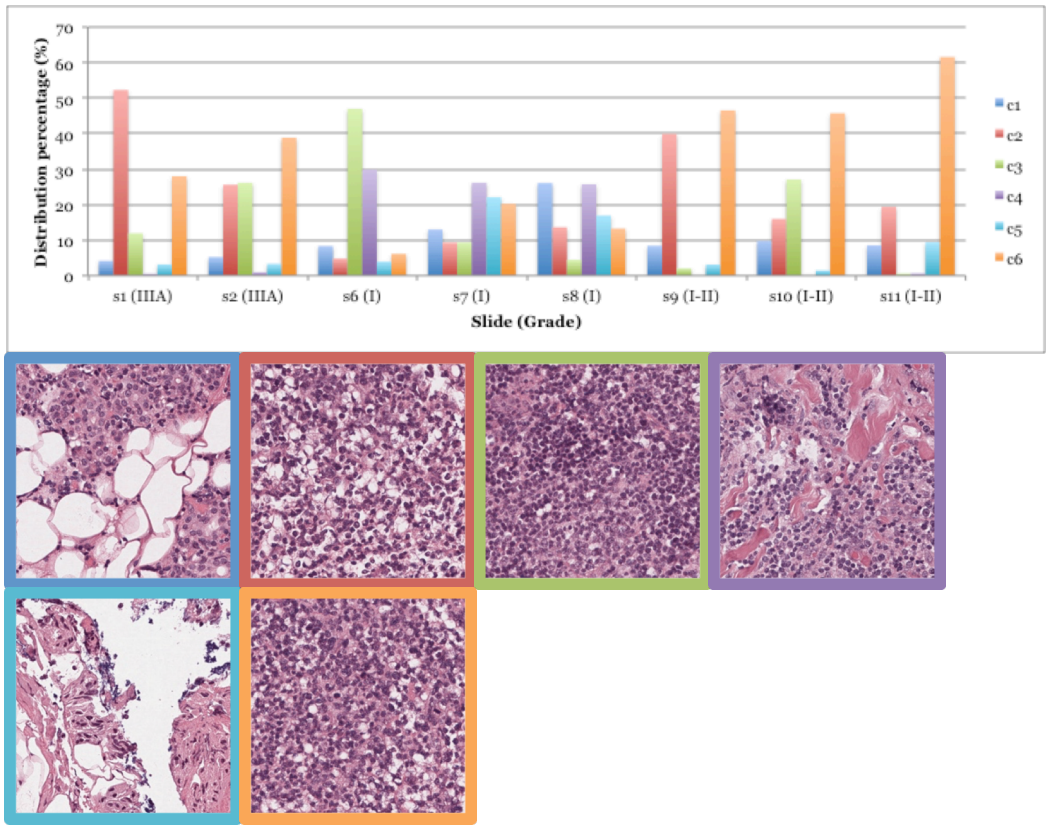


Figure 6.13 Cluster distribution (as percentage) per slide, K=6.

With the same strategy, we can examine the cluster distribution patterns revealed in Figure 6.14. However, with K=4, it is more difficult to differentiate slides between grades. One explanation is that four clusters of tile pattern are not specific enough to cover major variations among tiles and some subtle differences are now grouped together and are therefore overshadowed by prominent visual pattern categories.

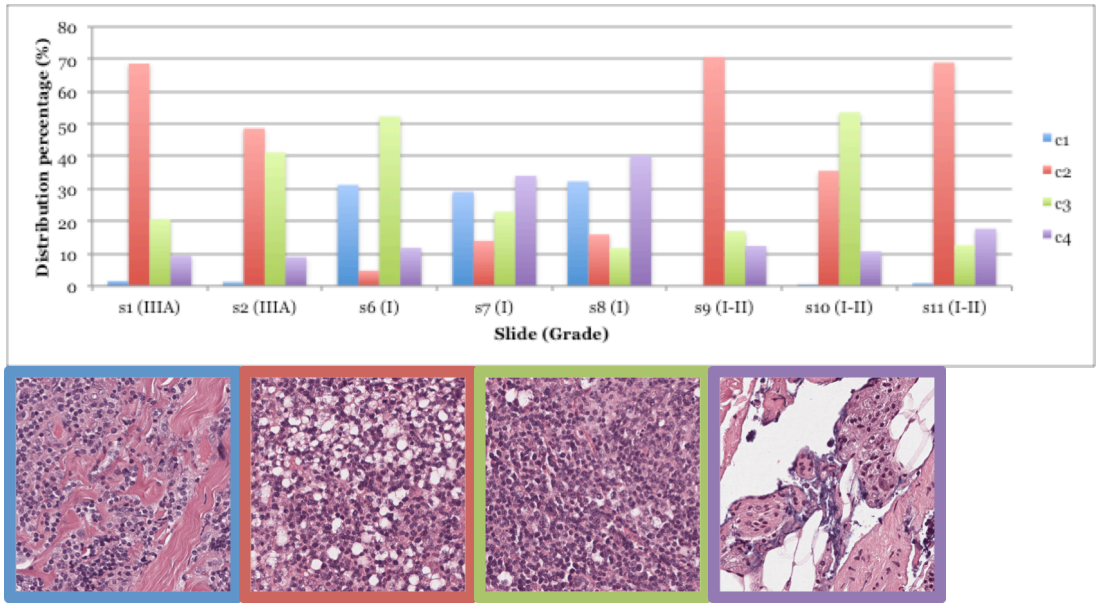


Figure 6.14 Cluster distribution percentage (%) per slide, K = 4.

As shown in Figure 6.15, as K value increases, clusters are further divided. When K increases from 4 to 6, clusters 2 and 3 (K=4) are now represented with three clusters (clusters 2, 4, and 6, K=6). When K increases from 6 to 8, six tile clusters now represent previously discovered four visual categories. The overlapping clusters indicate that finer details (appearance of fiber cells) are considered when differentiating between clusters. These observations point out the suggestion that choosing a single K is not always the ultimate goal. Instead, we should provide flexible options for end-users, in our case pathologists, to make judgments. They are the ultimate decision makers who control the resolution of cluster details. Choosing multiple K values could also provide a continuous discovering path for analyzing dynamic changes in visual differences.

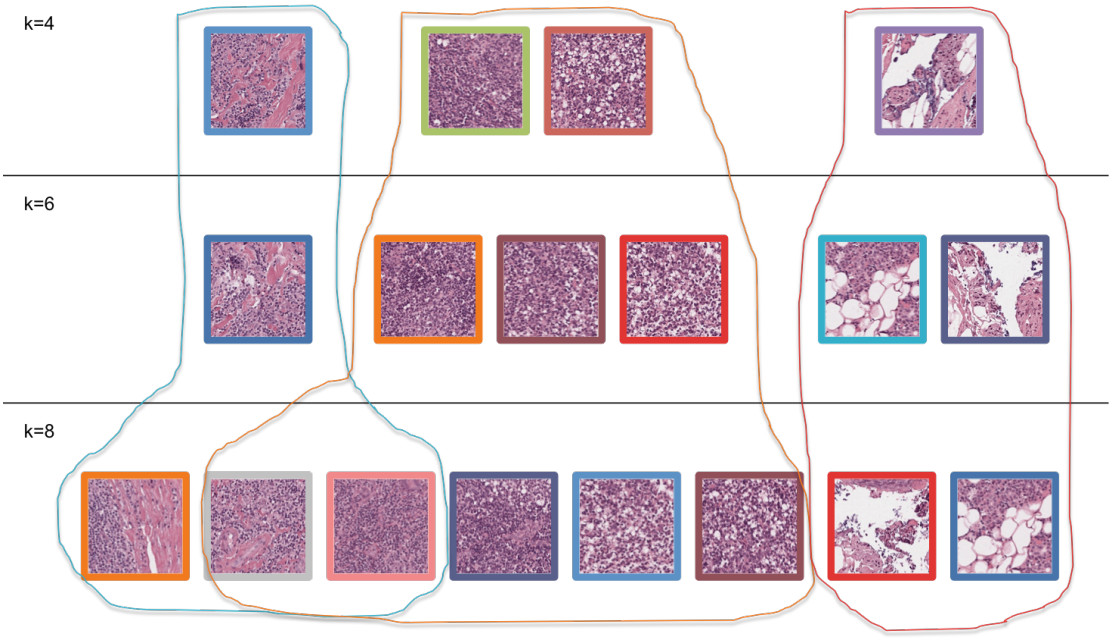


Figure 6.15 Visual categories discovered using different *K* values.

6.3.8 CBIR for Tiles

In addition to discovering visual categories, high-dimensional feature vectors can also help research to compare and retrieve tiles in a CBIR system. Figure 6.16 shows four CBIR cases using query tile images. With one query tile image, a list of tiles that are ranked based on their content similarity with the query image is presented.

- Query tile *q1* contains a mixture of lymphocytes and fatty tissue (white regions). From it retrieval list, *r1* and *r2* present similar pattern as *q1*, in the meantime, as similarity decreases, *r3* to *r5* show not only lymphocytes and fatty tissue regions but also glands that are also appearing to be white regions.
- Query tile *q2* and its retrieval result have the most consistent performance. All five tiles are representing dominant visual pattern of “starry sky”.

- Query tile *q3* retrieves a list of tiles that present both lymphocytes and fiber tissue regions.
- Query tile *q4* has a less advanced "starry sky" pattern. As returned from the retrieval, 4 out of 5 top-ranked tiles show similar visual pattern with an exception of *r4* that contains majority normal lymphocytes and less macrophages.

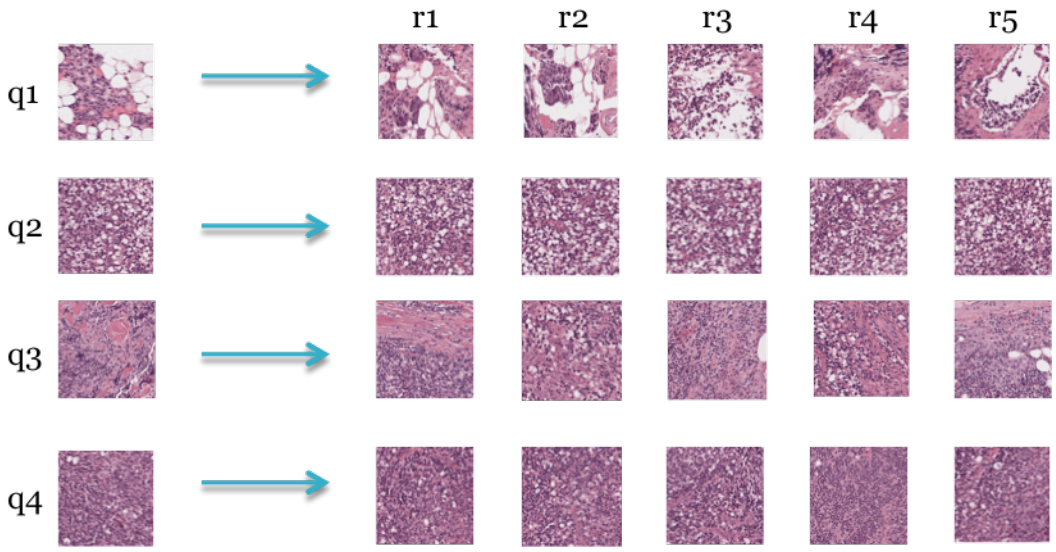


Figure 6.16 Content-based retrieval using tile samples. Results are ranked based on feature vector distances between query image and database images.

From this retrieval experiment, we demonstrate the capability of finding most visually similar tile images from a large collection of heterogeneous tiles. With this method, pathologists can submit a limited set of tiles of interest and ask the CBIR system to automatically retrieve a list of top-ranked most similar tiles based solely on their visual content. Accompanied with the discovered visual categories, we present a computer-assisted approach to study the gross visual patterns residing inside raw whole-slide images in for digital pathology domain experts.

6.4 Discussion

In this chapter, we presented our research works in pathology informatics using Big Data infrastructure. Particularly, we have processed and analyzed whole-slide images from patients with Follicular Lymphoma and Diffuse Large B-Cell Lymphoma with developed computer vision methods and constructed and executed with Big Data technologies, such as Hadoop, HDFS, and Apache Spark. We demonstrated the challenges in processing large-scale image as well as the complex content residing inside medical images. Our approaches have the following strengths in advancing large-scale microscopic image analysis. Whole-slide images are large in size and bear complex pathological content. We first divide them into smaller tiles based on the observation that conventional examinations of high power fields are done within a smaller window of viewing field. The localized visual patterns are comprised with rich pixel information and are extracted using a series of computer vision methods. Individual cells (cell nuclei) are identified with maximal preservation in morphology and used to construct cell graph as part of the process to extract visual content of tile images. The tiles, represented by a multi-dimensional feature vectors are studied to discover categories of visual patterns that are distributed throughout the whole slide images. The whole process was structured under the Big Data infrastructure. Specifically, we utilize Apache Spark as the main package handling distributed computing and the Hadoop Distributed File System (HDFS) to store data files across the cluster. Our scalability experiment shows that the Big Data setup has its merits in scalability especially for processing large-scale images, sometime billions of data instances, with efficiency and fault tolerance.

With that being said, the Big Data technologies bear some limitations and controversies as well [100, 101]. In the field of computer vision, images are often considered as a special case of data matrices. In some of the cases, image processing is performed as variations of matrix operation, such as image smoothing, convolution, and Fourier transformation. However, some of the advanced processing procedures do not necessarily fall under matrix operation. In our study, we use DMP operations to segment cells with moderate overlapping and blurriness. This operation has a great success in object segmentation, however, it was first developed as a recursive and therefore its execution time is not as efficient as some other approaches. Although in the following years since its first introduction, some variations were presented with improvement in computation efficiency [92, 95]. The parallelizable version of DMP operation was developed for execution on share-memory super computers with MPI mechanism. On the other hand, Spark, and Big Data technologies in general, work in a distributed computational structure where workers execute programs independently without messaging and communication among each other, except communication with the master node for overall process coordination. This difference hinders us from direct adoption of DMP and its parallel version into Big Data ecosystem.

In this work we demonstrate that, with moderate variations, distributed image processing can be achieved by shipping heavily specialized and complex image processing programs to individual workers for execution. Furthermore, the traffic of shipping data from master node to workers nodes is another aspect that needs to be carefully examined and treated. For digital pathology images, billions

of pixels are stored and are potentially useful for extracting image content. The intuition of distributing pixels onto workers nodes for in-memory computing is in fact impractical. That is the other reason that we only deliver necessary meta data information to workers and let them run heavy-duty image processing on their CPUs and save result imaged directly on to distributed file systems without shipping them back to the master node (driver program). The stress that is put on the driver program to gather all processed image data and save them to disk can be reaching over the limit as the driver program only runs on a single node.

In all, we recommend distributed image processing to be executed directly on workers nodes and leaving the coordination of different image patches (in our case the tiles) to be handled by driver program. Taking advantage of the ability of Big Data technologies to handle large-scale computation with distributed mechanism does not mean that every type of computation and data processing is necessarily in-memory and across the cluster.

CHAPTER SEVEN

CONCLUSIONS AND FUTURE WORK

In this chapter, we will review the major contributions presented in this dissertation as well as discuss some directions for future research based on the developed methods.

7.1 Conclusions

This dissertation centers around the development of a series of computational and informatics methods designed to assist large-scale image analysis, management, and retrieval, with an emphasis placed on biological and medical images. A summary of these contributions follows:

The aspect of web-based image management is first addressed in this dissertation by showcasing three real-world applications for domain experts. Specifically, a web platform (BioShapes.org) was developed as part of the multi-institute and multi-disciplinary collaboration project among biologists, computer scientists, mathematicians, and informaticians based on their common interest in understanding how biological shapes contribute to biological functionalities in a diverse selection of organisms. As part of the BioShapes web site, we developed a hierarchical image annotation and labeling tool for domain experts to quickly and accurately group and annotate large collection of mitochondria images based on their shape characteristics. This web-based tool has its back-end engine to first cluster visually similar mitochondria images using extracted visual features. Next, pre-clustered images are presented for manual review and adjustment. Visually similar images that are assigned in the same cluster can be presumably

annotated with the same label as their cluster example. Therefore, only a limited number of images need manual annotation. Once images are reliably annotated, they are again examined and assigned with multiple classes of morphological characteristics in shapes and sizes. This method is proven to be both efficient and accurate and helped biologists to reduce laborious examination so they can concentrate on the details analysis on a limited set of representative images. Such application can be easily adapted into other similar biological image analysis practices. For example, the Neotropical pollen and spore image database can also be benefited from efficient morphology annotation to improve the quality of taxonomy study. This research work points out the challenges as well as our solutions on complex annotation of biological images. A computational approach could ease the burden of what used to be laborious and potentially bias processes.

Two web-based dermatology image management and consultation systems were also introduced. FIRST Tele-Ichthyosis has been helping the global community of patients and dermatologists who are interested in the treatment and management of Ichthyosis (a rare skin disease) using our web-based tele-consultation system where medical cases along with patient images are uploaded to a secure webserver for domain experts to do online consultation and communication between experts and with case doctors. A study was conducted to understand the communication complexity and how online interactions between doctors and cases are constructed.

Mizzou Dermatology Image Database (MDID) is a web-based clinical image management and annotation system for the dermatology professionals at the University of Missouri Dermatology Department. We not only provided an

alternative image management application for better security and efficiency, but also went further to study the usability of such web-based image management system and how its users interact with it with salient usage patterns. System usage log data was first summarized using 2-round of mapping to transfer raw web browsing actions to meaningful tasks. Then sequences of tasks were studied using sequential pattern mining technique, SPADE, to discover usage pattern. A few recommendations for developing a better health IT system with high quality in usability and high adoption rate are presented. These two research works in dermatology domain can be extended into other medical domains that rely heavily on medical images. Our works demonstrate the importance of studying communications, interactions (both between human experts and between human users and health IT systems), adoptions of health IT applications in the real-world settings. Data mining techniques (i.e. social networking construction and sequential pattern mining) strengthened the studies by providing computational evidence to support our findings and in turn help us make recommendations for future health IT adoption practices.

Next, we presented the research works in visual content extraction in the fields of radiology, pathology, and palynology. Novel approaches were developed to handle object segmentation such as a multi-level follicle identification in whole-slide IHC images of follicular lymphoma cases; pathology-bearing regions in HRCT images of lung were analyzed and identified using modularized PC recognizers for different categories. We demonstrate how automatic parameter tuning would improve the overall quality of image analysis and retrieval. This not

only benefits the radiology domain but also is a generic approach for visual content extraction tasks for other imaging informatics domain.

Following visual content extraction, we then presented our works in content-based image retrieval for biological and medical images. To accommodate multi-class CBIR for HRCT images of lung, we developed a novel approach using the entropy of retrieval result to re-weight and re-rank the image to present a consolidated retrieval results. We also developed a web-based image retrieval system for Neotropical pollen and spore images. Both semantic-based and content-based image retrieval were provided using extracted morphological characteristics. These research works demonstrate that visual content extracted from raw images provides an alternative solution to traditional analyses of biological and medical images with capability of handling large-scale and complex image collections with efficiency and accuracy.

We conclude the presentation of our research works with a pathology image analysis system utilizing Big Data technologies. Due the large-scale image size and complexity, we first developed as series of image analysis method, including cDMP, to smartly identify cell nuclei from isolated H-stain images. Then we execute these heavy-duty computer vision programs under the Apache Spark computing cluster infrastructure. The efficiency, scalability, as well as performance in discovering pathologically meaningful visual categories were conducted and presented. We demonstrate the capability of analyzing raw medical images with billions of raw pixels bearing complex and high-level domain-specific knowledge, discovering underlying visual patterns, and assisting

pathologists to answer medically meaningful questions such as disease grading in our case study.

7.2 Future Work

7.2.1 Development of Methods for Other Imaging Domains

In both biology and medicine, there are likely a large number of visual characteristics that domain expert use in their daily practices to study various species, disease, morphologies, etc. A set of novel computer vision methods would facilitate the ever-increasing needs in high-throughput and large-scale analyses. The challenges are to collaborate with domain experts and understand the true nature of their "perceptual categories" and translate their needs and knowledge into useful computer vision tools.

7.2.2 In-Depth and Large-Scale Evaluations on Developed Methods

We acknowledge the limitations on some of our works in both the scalability and flexibility in handling biological and medical imaging data. We would like to dedicate great efforts on working with domain experts to validate our approaches on large-scale and more diverse collections of imaging data. An easy-to-use web-based application should be developed to facilitate expert labeling and validation processes to obtain high quality and possibly bigger volume of ground truth dataset.

7.2.3 Multi-Source Data Analytics Tools for Biomedical Imaging Informatics

For the current works, we are mainly dealing with a single source of data; whether they are images or text data. In the era of Big Data and ever-expanding Internet of things, we would expect various sources of related data to be collected

and waiting to be analyzed to make sense of scientific questions. One area is of particular interest. That is merging genetic, medical, social, and image information into a health informatics analysis system that would be able to handle biomedical Big Data and answer health-related question that may benefit a large population of patients as well as general society.

BIBLIOGRAPHY

1. Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *352*(1358), 1121-1127.

2. De Vos, K. J., & Sheetz, M. P. (2007). Visualization and quantification of mitochondrial dynamics in living animal cells. *Methods in cell biology*, *80*, 627-682.

3. Gonzalez, R. C., & Woods, R. (2002). Digital image processing. *Pearson Education*.

4. Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, (6), 610-621.

5. Luccheseyz, L., & Mitray, S. K. (2001). Color image segmentation: A state-of-the-art survey. *Proceedings of the Indian National Science Academy (INSA-A)*, *67*(2), 207-221.

6. Cheng, H. D., Jiang, X. H., Sun, Y., & Wang, J. (2001). Color image segmentation: advances and prospects. *Pattern recognition*, *34*(12), 2259-2281.

7. Pal, N. R., & Pal, S. K. (1993). A review on image segmentation techniques.*Pattern recognition*, *26*(9), 1277-1294.

8. Otsu, N. (1975). A threshold selection method from gray-level histograms.*Automatica*, *11*(285-296), 23-27.

9. Chang, S. G., Yu, B., & Vetterli, M. (2000). Spatially adaptive wavelet thresholding with context modeling for image denoising. *Image Processing, IEEE Transactions on*, *9*(9), 1522-1531.

10. Tcheslavski, G. V. (2010). Morphological Image Processing: Gray-scale morphology. *ELEN*, *4304*, 5365.

11. Vincent, L. (1993). Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. *Image Processing, IEEE Transactions on*, *2*(2), 176-201.

12. Lantuéjoul, C., & Maisonneuve, F. (1984). Geodesic methods in quantitative image analysis. *Pattern recognition*, *17*(2), 177-187.

13. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*, 1157-1182.

14. Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International journal of computer vision*, *7*(1), 11-32.

15. Hu, M. K. (1962). Visual pattern recognition by moment invariants.*information Theory, IRE Transactions on*, *8*(2), 179-187.

16. Zhang, D., & Lu, G. (2002). Generic Fourier descriptor for shape-based image retrieval. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on* (Vol. 1, pp. 425-428). IEEE.

17. Zhou, F., Feng, J. F., & Shi, Q. Y. (2001, October). Texture feature based on local Fourier transform. In *Image Processing, 2001. Proceedings. 2001 International Conference on* (Vol. 2, pp. 610-613). IEEE.

18. Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *JOSA A*, *2*(7), 1160-1169.

19. Chang, T., & Kuo, C. J. (1993). Texture analysis and classification with tree-structured wavelet transform. *Image Processing, IEEE Transactions on*, *2*(4), 429-441.

20. Lew, M. S., Sebe, N., Djeraba, C., & Jain, R. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *2*(1), 1-19.

21. Kunil, T. L. (1981). Pictorial data-base systems. *Computer*, (11), 13-21.

22. Müller, H., Michoux, N., Bandon, D., & Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *International journal of medical informatics*, *73*(1), 1-23.

23. Bishop, C. M. (2006). Pattern Recognition. *Machine Learning*.

24. Zheng, K., Padman, R., Johnson, M. P., & Diamond, H. S. (2009). An interface-driven analysis of user interactions with an electronic health records system. *Journal of the American Medical Informatics Association*, *16*(2), 228-237.

25. Hirsch, O., Szabo, E., Keller, H., Kramer, L., Krones, T., & Donner-Banzhoff, N. (2012). arriba-lib: Analyses of user interactions with an electronic library of decision aids on the basis of log data. *Informatics for Health and Social Care*,*37*(4), 264-276.

26. Horsky, J., McColgan, K., Pang, J. E., Melnikas, A. J., Linder, J. A., Schnipper, J. L., & Middleton, B. (2010). Complementary methods of

system usability evaluation: surveys and observations during software design and development cycles. *Journal of biomedical informatics*, *43*(5), 782-790.

27. Qiu, M., Lee, H. C., & Yang, G. (2012, May). Nanometer resolution tracking and modeling of bidirectional axonal cargo transport. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on* (pp. 992-995). IEEE.

28. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In*Kdd* (Vol. 96, No. 34, pp. 226-231).

29. Barb, A. S., & Shyu, C. R. (2010). Visual-semantic modeling in content-based geospatial information retrieval using associative mining techniques.*Geoscience and Remote Sensing Letters, IEEE*, *7*(1), 38-42.

30. http://www.americantelemed.org

31. Hicks, L. L., Boles, K. E., Hudson, S., Kling, B., Tracy, J., Mitchell, J., & Webb, W. (2003). Patient satisfaction with teledermatology services. *Journal of telemedicine and telecare*, *9*(1), 42-45.

32. www.firstskinfoundation.org

33. Agrawal, R., & Srikant, R. (1995, March). Mining sequential patterns. In *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*(pp. 3-14). IEEE.

34. Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine learning*, *42*(1-2), 31-60.

35. Traverse, A. (2007). Paleopalynology. Dordrecht, The Netherlands: Springer.

36. Faegri, K., Kaland, P. E., & Krzywinski, K. (1989). *Textbook of pollen analysis*(No. Ed. 4). John Wiley & Sons Ltd..

37. Birks, H. J., & Peglar, S. M. (1980). Identification of Picea pollen of Late Quaternary age in eastern North America: a numerical approach. *Canadian Journal of Botany*, *58*(19), 2043-2058.

38. Mander, L., & Punyasena, S. W. (2014). On the taxonomic resolution of pollen and spore records of Earth's vegetation. *International Journal of Plant Sciences*, *175*(8), 931-945.

39. Mander, L., Baker, S. J., Belcher, C. M., Haselhorst, D. S., Rodriguez, J., Thorn, J. L., ... & Punyasena, S. W. (2014). Accuracy and consistency of

grass pollen identification by human analysts using electron micrographs of surface ornamentation. *Applications in plant sciences*, *2*(8).

40. Bush, M. B., & Weng, C. (2006). Introducing a new (freeware) tool for palynology. *Journal of Biogeography, 34*(3): 377-380.

41. www.paldat.org

42. www.neotomadb.org

43. Grimm, E. C., Keltner, J., Cheddadi, R., Hicks, S., Lézine, A.-M., Berrio, J. C., & Williams, J. W. (2013). Pollen databases and their application. In Elias, S. A. and C. J. Mock [eds.]. *Encyclopedia of Quaternary Science*, 831-83. Elsevier.

44. Holt, K., Allen, G., Hodgson, R., Marsland, S., & Flenley, J. (2011). Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. *Review of Palaeobotany and Palynology*,*167*(3), 175-183.

45. Holt, K. A., & Bennett, K. D. (2014). Principles and methods for automated palynology. *New Phytologist*, *203*(3), 735-742.

46. Quiroz, L., & Jaramillo, C. (2010). Stratigraphy and sedimentary environments of Miocene shallow to marginal marine deposits in the Urumaco Trough, Falcon Basin, western Venezuela. *Urumaco and Venezuelan Paleontology*, 153-172.

47. Aguilera, O. A., & Carlini, A. A. (Eds.). (2010). Urumaco and Venezuelan Paleontology: the fossil record of the Northern Neotropics. Indiana University Press.

48. Jaramillo, C., & Rueda, M. (2008). A morphological electronic database of Cretaceous-Tertiary fossil pollen and spores from northern South America.*Colombian Petroleum Institute & Smithsonian Tropical Research Institute.*

49. Ibañez, L., Schroeder, W., Ng, L., & Cates, J. (2003). The ITK software guide.

50. Abràmoff, M. D., Magalhães, P. J., & Ram, S. J. (2004). Image processing with ImageJ. *Biophotonics international*, *11*(7), 36-42.

51. OpenCV, L. (2008). Computer vision with the OpenCV library. *Gary Bradski & Adrian Kaebler-O'Reilly*.

52. Pham, D. L., Xu, C., & Prince, J. L. (2000). Current methods in medical image segmentation 1. *Annual review of biomedical engineering*, *2*(1), 315-337.

53. Armato, S. G., & MacMahon, H. (2003, June). Automated lung segmentation and computer-aided diagnosis for thoracic CT scans. In *International Congress Series* (Vol. 1256, pp. 977-982). Elsevier.

54. Russ, J. C. (2015). *The image processing handbook*. CRC press.

55. Ohta, Y. I., Kanade, T., & Sakai, T. (1980). Color information for region segmentation. *Computer graphics and image processing*, *13*(3), 222-241.

56. Kirkpatrick, S. (1984). Optimization by simulated annealing: Quantitative studies. *Journal of statistical physics*, *34*(5-6), 975-986.

57. Han, J. G., & Shyu, C. R. (2010). Improving retrieval performance in medical image databases using simulated annealing. In *AMIA Annual Symposium Proceedings* (Vol. 2010, p. 276). American Medical Informatics Association.

58. Vincent, L., & Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6), 583-598.

59. Punt, W., Hoen, P. P., Blackmore, S., Nilsson, S., & Le Thomas, A. (2007). Glossary of pollen and spore terminology. *Review of Palaeobotany and Palynology*, *143*(1), 1-81.

60. Barb, A., & Kilicay-Ergin, N. (2013). Genetic optimization for associative semantic ranking models of satellite images by land cover. *ISPRS International Journal of Geo-Information*, *2*(2), 531-552.

61. Krishnapuram, R., & Keller, J. M. (1993). A possibilistic approach to clustering. *Fuzzy Systems, IEEE Transactions on*, *1*(2), 98-110.

62. Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, No. 2, pp. 1137-1145).

63. Samsi, S., Lozanski, G., Shanarah, A., Krishanmurthy, A. K., & Gurcan, M. N. (2010). Detection of follicles from IHC-stained slides of follicular lymphoma using iterative watershed. *Biomedical Engineering, IEEE Transactions on*, *57*(10), 2609-2612.

64. Doyle, S., Madabhushi, A., Feldman, M., & Tomaszeweski, J. (2006). A boosting cascade for automated detection of prostate cancer from digitized histology. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006* (pp. 504-511). Springer Berlin Heidelberg.

65. www.aperio.com

66. Zijdenbos, A. P., Dawant, B. M., Margolin, R. A., & Palmer, A. C. (1994). Morphometric analysis of white matter lesions in MR images: method and validation. *Medical Imaging, IEEE Transactions on*, *13*(4), 716-724.

67. Hersh, W., Müller, H., & Kalpathy-Cramer, J. (2009). The ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging*, *22*(6), 648-655.

68. Chong, E. K., & Zak, S. H. (2013). *An introduction to optimization* (Vol. 76). John Wiley & Sons.

69. Raittinen, H., & Kaski, K. (1990). Image deconvolution with simulated annealing method. *Physica Scripta*, *1990*(T33), 126.

70. Smeulders, A. W., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *22*(12), 1349-1380.

71. Niblack, C. W., Barber, R., Equitz, W., Flickner, M. D., Glasman, E. H., Petkovic, D., ... & Taubin, G. (1993, April). QBIC project: querying images by content, using color, texture, and shape. In *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology* (pp. 173-187). International Society for Optics and Photonics.

72. Pentland, A. P., Picard, R. W., & Scarloff, S. (1994, April). Photobook: Tools for content-based manipulation of image databases. In *IS&T/SPIE 1994 International Symposium on Electronic Imaging: Science and Technology*(pp. 34-47). International Society for Optics and Photonics.

73. Smith, J. R., & Chang, S. F. (1997, February). VisualSEEk: a fully automated content-based image query system. In *Proceedings of the fourth ACM international conference on Multimedia* (pp. 87-98). ACM.

74. , C. R., Klaric, M., Scott, G. J., Barb, A. S., Davis, C. H., & Palaniappan, K. (2007). GeoIRIS: Geospatial information retrieval and indexing system—Content mining, semantics modeling, and complex queries. *Geoscience and Remote Sensing, IEEE Transactions on*, *45*(4), 839-852.

75. Csillaghy, A., Hinterberger, H., & Benz, A. O. (2000). Content-based image retrieval in astronomy. *Information retrieval*, *3*(3), 229-241.

76. Shyu, C. R., Chi, P. H., Scott, G., & Xu, D. (2004). ProteinDBS: a real-time retrieval system for protein structure comparison. *Nucleic Acids Research*,*32*(suppl 2), W572-W575.

77. Brenner, D. J., & Hall, E. J. (2007). Computed tomography—an increasing source of radiation exposure. *New England Journal of Medicine*, *357*(22), 2277-2284.

78. Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, *22*(2), 207-216.

79. Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463). New York: ACM press.

80. Ciaccia, P., Patella, M., & Zezula, P. (1997). DEIS-CSITE-CNR. In*Proceedings of the... International Conference on Very Large Data Bases*(Vol. 23, p. 426). Morgan Kaufmann Pub.

81. Scott, G., & Shyu, C. R. (2007). Knowledge-driven multidimensional indexing structure for biomedical media database retrieval. *Information Technology in Biomedicine, IEEE Transactions on*, *11*(3), 320-331.

82. Punyasena, S. W., Tcheng, D. K., Wesseln, C., & Mueller, P. G. (2012). Classifying black and white spruce pollen using layered machine learning.*New Phytologist*, *196*(3), 937-944.

83. Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 3.

84. Murdoch, T. B., & Detsky, A. S. (2013). The inevitable application of big data to health care. *Jama*, *309*(13), 1351-1352.

85. Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS quarterly*, *36*(4), 1165-1188.

86. Amatriain, X. (2013). Mining large streams of user data for personalized recommendations. *ACM SIGKDD Explorations Newsletter*, *14*(2), 37-48.

87. Lin, J., & Ryaboy, D. (2013). Scaling big data mining infrastructure: the twitter experience. *ACM SIGKDD Explorations Newsletter*, *14*(2), 6-19.

88. Ruifrok, A. C., & Johnston, D. A. (2001). Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology/the International Academy of Cytology [and] American Society of Cytology, 23(4), 291-299.

89. Veta, M., van Diest, P. J., Kornegoor, R., Huisman, A., Viergever, M. A., & Pluim, J. P. (2013). Automatic nuclei segmentation in H&E stained breast cancer histopathology images. PloS one, 8(7), e70221.

90. Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Guan, X., ... & Thomas, N. E. (2009, June). A Method for Normalizing Histology Slides for Quantitative Analysis. In *ISBI* (Vol. 9, pp. 1107-1110).

91. Jette, M., & Grondona, M. (2003). SLURM: Simple Linux Utility for Resource Management. Proceedings of ClusterWorld Conference and Expo. San Jose, California.

92. Klaric, M., Scott, G., Shyu, C. R., & Davis, C. (2005, July). Automated object extraction through simplification of the differential morphological profile for high-resolution satellite imagery. In *Geoscience and Remote Sensing Symposium, 2005. IGARSS'05. Proceedings. 2005 IEEE International* (Vol. 2, pp. 1265-1268). IEEE.

93. Ouzounis, G. K., Soille, P., & Pesaresi, M. (2011, April). Rubble detection from VHR aerial imagery data using differential morphological profiles. In*34th Int. Symp. Remote Sensing of the Environment*.

94. Dalla Mura, M., Benediktsson, J. A., Waske, B., & Bruzzone, L. (2010). Morphological attribute profiles for the analysis of very high resolution images. *Geoscience and Remote Sensing, IEEE Transactions on*, *48*(10), 3747-3762.

95. Pesaresi, M., & Benediktsson, J. A. (2001). A new approach for the morphological segmentation of high-resolution satellite imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, *39*(2), 309-320.

96. Bradley, D., & Roth, G. (2007). Adaptive thresholding using the integral image. *Journal of graphics, gpu, and game tools*, *12*(2), 13-21.

97. Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *24*(5), 603-619.

98. Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219(1), 103-119.

99. Bahmani, B., Moseley, B., Vattani, A., Kumar, R., & Vassilvitskii, S. (2012). Scalable k-means++. *Proceedings of the VLDB Endowment*, *5*(7), 622-633.

100. Fan, W., & Bifet, A. (2013). Mining big data: current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter*, *14*(2), 1-5.

101. Che, D., Safran, M., & Peng, Z. (2013, April). From big data to big data mining: challenges, issues, and opportunities. In *Database Systems for Advanced Applications* (pp. 1-15). Springer Berlin Heidelberg.

VITA

Jing Han received her PhD degree in Health Informatics from the University of Missouri in 2016. Previously she received her BS degree from Department of Electrical Engineering, Xi'an Jiaotong University, China.

Jing's PhD research interest consists of computer vision, machine learning, pattern recognition, information retrieval, web-based health IT applications, and broad topics in health informatics. Since 2008, during her PhD training as a research assistant, she has participated in multiple multi-disciplinary research activities. She has broad experience in working with professionals in the fields of biology, palynology, radiology, cardiology, pathology, dermatology, and non-profit organization. She also served as the Treasurer of MUII Graduate Student Association from 2010 to 2011. Besides research duties, she also served as a Graduate Instructor of MU Howard Hughes Medical Institute Undergraduate Summer Biomedical Informatics Institute 2013, the teaching assistant for course MUII 7001 (Introduction to Informatics) in 2013, and the co-instructor for course MUII 7005 (Introduction to Bioinformatics) in 2013. She was one of the student members who conceived the idea of hosting Missouri Informatics Day in 2011 (later expanded to Missouri Informatics Symposium) and also served in the committee for MIS 2011 and MIS 2013. She is currently a Data Scientist at College of Veterinary Medicine, Mississippi State University. She is also a student member in AMIA and Botany.