

RICE UNIVERSITY

**Automated Detection and Differential Diagnosis of
Non Small Cell Lung Carcinoma Cell Types Using
Label-Free Molecular Vibrational Imaging**

by

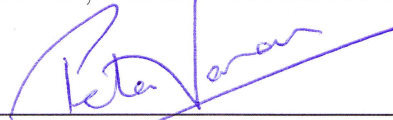
Ahmad A. Hammoudi

A THESIS SUBMITTED

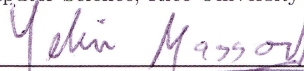
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Science

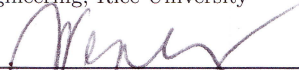
APPROVED, THESIS COMMITTEE:



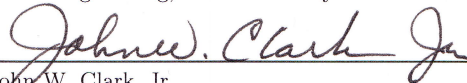
Dr. Peter Varman
Professor of Electrical and Computer Engineering and
Computer Science, Rice University




Dr. Yehia Massoud
Chair, Department of Electrical and Computer
Engineering
Wallace R Bunn Endowed Chair in Telecommunications,
The University of Alabama at Birmingham
Adjunct Professor of Electrical and Computer
Engineering, Rice University



Dr. Stephen T.C. Wong
Chair, Department of Systems Medicine and
Bioengineering
John S Dunn Distinguished Endowed Chair, The
Methodist Hospital Research Institute
Professor of Radiology, Pathology, and Laboratory
Medicine, Cornell University
Adjunct Professor of Bioengineering and Electrical and
Computer Engineering, Rice University



Dr. John W. Clark, Jr.
Professor of Electrical and Computer Engineering and
Bioengineering, Rice University



Dr. Behnaam Aazhang
Chair, Department of Electrical and Computer
Engineering
J.S. Abercrombie Professor of Electrical and Computer
Engineering, Rice University

Houston, Texas

April, 2012

RICE UNIVERSITY

**Automated Detection and Differential Diagnosis of
Non Small Cell Lung Carcinoma Cell Types Using
Label-Free Molecular Vibrational Imaging**

by

Ahmad A. Hammoudi

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
Master of Science

APPROVED, THESIS COMMITTEE:

Dr. Peter Varman
Professor of Electrical and Computer Engineering and
Computer Science, Rice University

Dr. Yehia Massoud
Chair, Department of Electrical and Computer
Engineering
Wallace R Bunn Endowed Chair in Telecommunications,
The University of Alabama at Birmingham
Adjunct Professor of Electrical and Computer
Engineering, Rice University

Dr. Stephen T.C. Wong
Chair, Department of Systems Medicine and
Bioengineering
John S Dunn Distinguished Endowed Chair, The
Methodist Hospital Research Institute
Professor of Radiology, Pathology, and Laboratory
Medicine, Cornell University
Adjunct Professor of Bioengineering and Electrical and
Computer Engineering, Rice University

Dr. John W. Clark, Jr.
Professor of Electrical and Computer Engineering and
Bioengineering, Rice University

Dr. Behnaam Aazhang
Chair, Department of Electrical and Computer
Engineering
J.S. Abercrombie Professor of Electrical and Computer
Engineering, Rice University

Houston, Texas

April, 2012

*To my family,
the reason I stand where I do*

ABSTRACT

Automated Detection and Differential Diagnosis of Non Small Cell Lung Carcinoma
Cell Types Using Label-Free Molecular Vibrational Imaging

by

Ahmad A. Hammoudi

Advances in targeted therapy hold the premise for the delivery of more effective treatments to lung cancer patients, given the ability to diagnose and identify patient specific cell types, *small cell carcinoma*, *adenocarcinoma*, or *squamous cell carcinoma*. Label free optical imaging techniques like the *Coherent Anti-Stokes Raman Scattering microscopy* (CARS) can provide physicians with minimally invasive access to tumors and allow diagnosis and sub-typing. Exploiting CARS requires developing data analysis methods that can rapidly and accurately analyze the new types of data they provide. In this study we designed an image processing framework that automatically and accurately, detects cancer cells in two and three dimensional CARS images. Moreover, we built upon this capability with new approaches to analyzing the segmented data, that provided significant information about the cancerous tissue that allowed for the automatic differential classification of non-small cell lung carcinoma cell types, overcoming the shortcomings of previous such approaches.

Acknowledgments

No thanks are enough for my co-advisors Dr. Yehia Massoud, and Dr. Stephen Wong. The energy and enthusiasm that Yehia has, and so effortlessly transfers onto his students, is truly an inspiration. He is a great mentor on all matters academic, and personal alike. Dr. Wong through allowing me the privilege to work in his lab at The Methodist Hospital, willing to always provide any kind of support needed, be it financial, research advice, or professional advice, has truly made my stint in his lab, and my academic progress at Rice an easy and fun experience.

Many Thanks to my committee members Dr. Peter Varman, Dr. John Clark, and Dr. Behnaam Aazhang, for promptly taking the time to serve on the committee and oversee the defense.

Warm thanks go to all the members of the Systems Medicine and Bioengineering department at The Methodist Hospital Research Institute, and particularly to Liang Gao, Fuhai Li, Zhiyong Wang, and Lysa Donalson for all the help, data, invaluable discussions, and support they provided, and for enriching my experience through sharing theirs.

For the very special friends who always make life a little easier, David, Omar, Houssam, Imad, Khalid, Rayan, Ziad, Hussein, Peter, Hussain, Ahmed, Murtada, Yaser, Johnny, Karina, and especially Alda. I love you all so deeply, and I am so lucky to have you in my life.

Last but not least, I would not be anywhere close to where I am right now, if it wasn't for the love and support of my parents, Ali and Zakia. Very warm thanks to

My sisters and their husbands, Samar, Dalal, Said, and Abdulrahaman, for the joy they bring to my life, the great friends and greater family that they are. I love you.

Contents

Abstract	iv
Acknowledgments	v
List of Illustrations	x
List of Tables	xiii
1 Introduction	1
1.1 Lung Cancer Diagnosis and Treatment: The big Picture	1
1.2 Previous work	3
1.2.1 Label Free Microscopy with Coherent Anti-Stokes Raman Scattering	3
1.2.2 CARS imaging for Differential Diagnosis of Lung Carcinoma Cell Types	4
1.2.3 Limitations	6
1.3 Goals and a Forward for the Thesis	7
2 Two Dimensional Automatic Segmentation of Cell nuclei	11
2.1 The CARS Image Segmentation Problem	12
2.1.1 CARS Images and Standard Segmentation Approaches	12
2.1.2 Hierarchical Segmentation	13
2.1.3 Context Based Hierarchical Segmentation	13

2.2	Local Clustering in CARS Images	15
2.2.1	A Primer to Simple Linear Iterative Clustering Based Superpixels	16
2.2.2	Experimental Considerations of Superpixel Segmentation in CARS images of Various Lung Carcinoma Subtypes	20
2.3	Using Superpixel Context Information to Detect Nuclei in CARS Images	27
2.3.1	Defining Context Descriptors	28
2.3.2	Nuclear Superpixel Identification with an ANN	30
2.3.3	Results	34
2.4	Discussion	37
3	Three Dimensional Automatic Segmentation of Cell nu- clei	38
3.1	3D Segmentation: Why? And How?	38
3.1.1	introduction	38
3.1.2	The 3D CARS Image Segmentation Problem	42
3.2	2D to 3D: Superpixels to Supervoxels	44
3.3	Nuclei Detection From Supervoxels	48
3.3.1	Cluster Describing Features	48
3.3.2	Nuclei Detection with Semi-Supervised Learning Machines . .	53
3.4	Results and Discussion	57
3.4.1	3D Segmentation and Detection Results	57
3.4.2	Discussion	61
4	Automatic Differential Diagnosis of NSCLC Cell Types	63
4.1	Introduction	63

4.2	Ellipsoid Fitting	65
4.3	Designing Pathologically Relevant Features	69
4.3.1	Single Nucleus Features	70
4.3.2	Whole Tissue Features	71
4.4	Differential Classification	74
4.5	Results	79
4.6	Discussion	82
5	Conclusions and Future Work	86
	Bibliography	88

Illustrations

1.1	CARS images and corresponding H & E images, of healthy tissue (A, B), adenocarcinoma(C,D), and squamous cell carcinoma (E,F)	8
1.2	Overview of a complete CARS based automatic lung carcinoma diagnostic platform	10
2.1	The hierarchical approach used for segmentation and detection of cellular nuclei in CARS images	14
2.2	Superpixels generated in a CARS image of an Adenocarcinoma sample. Red markers represent the boundaries of superpixels. It is notable how well the superpixels adhere to image objects, namely nuclei	16
2.3	Illustration of potential data point-to-cluster assignment.	20
2.4	Design space exploration: effect of varying m and k on the generated superpixels	22
2.5	Convergence of SLIC clustering in a CARS image	23
2.6	Superpixels generated in a CARS image of a Small cell carcinoma sample. Red markers represent the boundaries of superpixels.	24
2.7	Superpixels generated in a CARS image of an Squamous cell carcinoma sample. Red markers represent the boundaries of superpixels	25
2.8	Gaussian filter used to de-noise CARS images	26
2.9	Illustration of ray descriptors (A) and Ellipsoidal goodness of fit(B) .	29

2.10	Schematic representation of a neuron	31
2.11	Schematic representation of interconnected neurons in a multilayer neural network	32
2.12	Segmentation and detection example: adenocarcinoma sample	34
2.13	Segmentation and detection example: small cell carcinoma sample . .	35
2.14	Segmentation and detection example: squamous carcinoma sample . .	36
3.1	A 3D CARS image stack of a lung tissue containing adenocarcinoma cells in its raw form	40
3.2	A 3D cross sectional representation of a CARS volume	41
3.3	A close up of a 3D cross sectional representation of a CARS volume, with nuclei visible in all 3 dimensions	42
3.4	Consecutive slices through a CARS volume showing a single nucleus being enclosed in a single supervoxel across all slices	47
3.5	A depiction of a general burst of rays in 3D space	51
3.6	Ray bursts inside a supervoxel (A). Intersection of rays with the supervoxel boundaries (B)	52
3.7	A segmentation and detection demonstration in s 3D CARS volume, with a rendering of the segmented and detected cell nuclei	58
3.8	Consecutive slices through a CARS volume - after classification - showing a nucleus being enclosed in a single supervoxel across all slices	58
3.9	Comparing supervoxel nuclei detection results from our method to other nuclei detection methodologies	60
4.1	A rendering of a full volume of segmented cell nuclei	66
4.2	A single nucleus (Blue), and the corresponding MVEE (Mesh)	68

4.3	MVEEs for all nuclei in a full CARS volume	69
4.4	An ellipsoid with its 3 direction vectors shown	71
4.5	Delaunay triangulation of nuclei in a CARS image	72
4.6	Statistics of single nucleus derived features - 2D vs. 3D	76
4.7	Statistics of tissue derived features - 2D vs. 3D	77
4.8	Differential diagnosis results from 2D data	80
4.9	Differential diagnosis results from 3D data	81

Tables

2.1	Validation scores for the segmentation and detection framework in all lung carcinoma subtypes	37
3.1	A quantitative comparison between graph transduction and SVM quality of supervoxel nuclei detection	61
4.1	Differential diagnosis accuracies from 2D data	81
4.2	Differential diagnosis accuracies from 3D data	81

*“If you torture the data long enough,
it will confess”*

- Ronald Coase

Chapter 1

Introduction

1.1 Lung Cancer Diagnosis and Treatment: The Big Picture

Lung carcinoma is the most prevalent type of cancer in the world responsible for more deaths than any other type of cancer [1]. It has been, for over 50 years, perceived as a relentlessly progressive, fatal disease. It is the primary cause of cancer deaths in the United States with 222,500 new cases of lung cancer and 157,300 lung cancer deaths reported in 2010 [2], with five-year survival rates less than 18% [1,3]. Globally, these rates drop to 6-14% for men and 7-18% for women [1,3]. These dismal mortality rates are caused first, by the lack of non-invasive or minimally invasive methods of early detection and diagnosis. And second, by the inability to develop effective targeted therapy, caused mainly by the inability to accurately determine which lung carcinoma cell types are present in diagnosed patients.

While early detection has attracted major research effort [4, 5], less than 1% of patients can be diagnosed at an early stage [6]. Tissue biopsy, a highly invasive method, is frequently needed as a follow-up test for definitive diagnosis, following pulmonary examination using computed tomography (CT) or magnetic resonance imaging (MRI). Nonetheless, it remains difficult to precisely locate the the site of small

lesions for obtaining the required samples [6]. Some patients will need to undergo re-biopsy, resulting in increased risks to patients, higher testing and treatment costs, and delays in diagnosis and treatment. So given the risks and cost of lung biopsy, it would be beneficial to develop an imaging strategy that can provide real time images and an accompanying computational diagnostic analysis platform of the biopsied site, relieving the requirements of accurately locating small lesions, limiting damage to lung tissue, diagnosing lung cancer in vivo, and providing diagnostic yield comparable to existing biopsy methods.

Moreover, pathologists have only achieved minor successes in differentiating the various cell types of lung carcinoma, namely *adenocarcinoma*, *squamous cell carcinoma*, and *small cell carcinoma*. In addition differentiating small cell from non-small cell carcinoma is a bigger challenge to pathologists and one that is crucial to developing targeted therapies [7]. This has led to categorizing the first two cell types with the term non-small cell lung carcinoma (NSCLC), to reflect this difficulty [7].

As such our work is framed by, and builds upon two broad ideas.

- Developing a minimally invasive automated diagnostic platform, that can rapidly and efficiently collect, process and analyze data from lung tissue, to locate and diagnose cancer cells.
- Developing an effective differential diagnosis strategy that can automatically provide information on the specific carcinoma cell types present in a tumor, to aid in designing and delivering optimized targeted therapy to patients.

1.2 Previous work

In subsection 1.2.1 we provide a brief overview of an imaging technique that holds great potential for use in diagnosis of lung carcinoma, followed by an overview of the most recent achievements in its application to lung cancer diagnosis in subsection 1.2.2. In subsection 1.2.3 we describe the shortcomings of the current state of the art in achieving the full potential of utilizing CARS in enhancing lung carcinoma diagnosis and differential classification.

1.2.1 Label Free Microscopy with Coherent Anti Stokes Raman Scattering

The coherent anti-Stokes Raman scattering (CARS), an optical imaging technique [8], holds great promise for diagnostic applications. It captures intrinsic molecular vibrations to create optical contrast with sub-micron level spatial resolution, as well as video-speed imaging rate [9]. In the CARS process three laser beams, a pump beam ω_p , a Stokes beam ω_s and a probe beam ω'_p interact with tissue samples through a four-wave mixing process [10]. When the frequency difference, $\omega_p - \omega_s$ (beat frequency), is in resonance with a molecular eigenvibration of CH chemical bonds in the tissue, an enhanced signal at the anti-Stokes frequency, $\omega_{as} = \omega_p - \omega_s + \omega'_p$, is generated [11], this signal is captured by a microscope to generate a CARS image.

A major advantage of CARS is that the signal yield is much higher, typically several orders of magnitude, than the signal yield obtained through the conventional

spontaneous Raman scattering process [12]. As a result, this imaging technique has been used to visualize different tissue structures, e.g. skin [12], lungs, kidney, and prostate [9]. In addition, major advantages of CARS include that it is a label-free, since it images intrinsic molecular vibrations, and it allows for 3D tissue sectioning [13, 14]. As such significant applications of CARS microscopy imaging in differential diagnosis of breast cancer and lung cancer have been reported [13, 14].

1.2.2 CARS imaging for Differential Diagnosis of Lung Carcinoma Cell Types

There have been some recent breakthroughs in the treatment of lung cancer, mainly through the use of targeted therapies, that promise to radically enhance survival rates of lung cancer patients [7]. However, the advent of molecularly targeted therapies makes identification of the various histologic cell types and subtypes of lung cancer a critical requirement. For example, adenocarcinoma patients should be tested for epidermal growth factor receptor (EGFR) mutations as an indication of responsiveness to EGFR tyrosine kinase inhibitor [15–17]. In addition, an exclusion of a squamous cell carcinoma diagnosis is required for NSCLC patients prior to treatment with bevacizumab because of potential life-threatening hemorrhage [18, 19]. As a result, additional molecular tests are frequently necessary for reaching a definitive diagnosis for targeted therapies [20, 21].

However, most lung carcinomas are not resected and are diagnosed and classified

using small biopsies or cytology specimens [22, 23]. Hematoxylin and eosin (H&E) staining of tissue sections, which is currently the gold standard for histologic diagnosis, requires hours to days for tissue transfer, processing, sectioning, and staining, and is still often incapable of differentiating NSCLC cell types. Cytology results are typically faster, but the tissue material is even more limited in volume, and reliably separating adenocarcinoma from squamous cell carcinoma is sometimes impossible. Finally, immunochemistry may be useful in helping to make this distinction, but this method adds more time to the diagnostic process and consumes tumor cells.

Therefore, the ability to rapidly recognize different cell types and subtypes of lung cancer with minimal tissue consumption, will not only facilitate the diagnostic process, but also enable maximum preservation of tissue samples for subsequent molecular testing for targeted therapy [21]. Given the risks and cost of lung biopsy, it would therefore be beneficial to develop techniques that enable fast examination of excised biopsy samples as a preliminary test for separating general cell types of lung cancer before a follow-up molecular analysis, with the aim of reducing the number of required biopsies and providing equal or greater accuracy relative to existing testing methods.

As a label-free imaging technique, CARS microscopy [8, 24] holds great potential for this type of diagnostic application by significantly minimizing sampling error and realizing maximum preservation of specimens for follow-up molecular tests. Nevertheless, to realize the diagnostic value of CARS, quantitative information must be

extracted and analyzed from the digital CARS images in order to meet the rigorous evaluation criteria required for objective diagnosis.

Accordingly, a pattern recognition and classification strategy that integrates CARS imaging with quantitative image computing techniques for cancer diagnosis has been reported [13,14,25–27]. This strategy was based on training classifiers with a series of pathologically-related morphological features, thus providing meaningful diagnostic information with reproducible results. The developed platform has been tested using a number of disease models, including lung, breast and prostate cancers [13,14,25–27]. Specifically to the case of lung carcinoma, previously developed CARS image analysis techniques have been able to computationally differentiate healthy from tumor tissue, benign from malignant tumors, and even small cell from non small cell lung carcinomas [13].

1.2.3 Limitations

In spite of this progress, two factors still hinder the full exploitation of the benefits offered by CARS microscopy to the advancement of diagnosis and treatment approaches.

First, image analysis methods that have been previously utilized in detecting cancer cells in CARS images are still far from being fully automatic [13,28]. They require an expert biologist to select each cell in a large data set of tissue images in order to utilize in the developed data analysis strategies that perform diagnosis.

Second, previous work has thus far shown limited accuracy - around 70 to 75% [25] - for separation of adenocarcinoma from squamous cell carcinoma, two major cell types of NSCLC. This difficulty is not surprising, and it corresponds with the clinical difficulty in differentiating these two cell types using morphology alone [29]. However, as discussed above, definitive diagnosis of these cell types is crucial to the advancement of targeted therapies [30].

1.3 Goals and a Forward for the Thesis

Building upon the previous work, with the goals stated in section 1.1 in mind, we aim to tackle the shortcomings still hindering the use of CARS to develop a minimally invasive strategy for detection and differential diagnosis of lung carcinoma. As described above, what is lacking to complete such a strategy is the ability to: And the ability to

- (a) Automatically analyze CARS images for cell segmentation, and information extraction.
- (b) Automatically differentiate the different NSCLC cell types, adenocarcinoma and squamous cell carcinoma, figure 1.1.

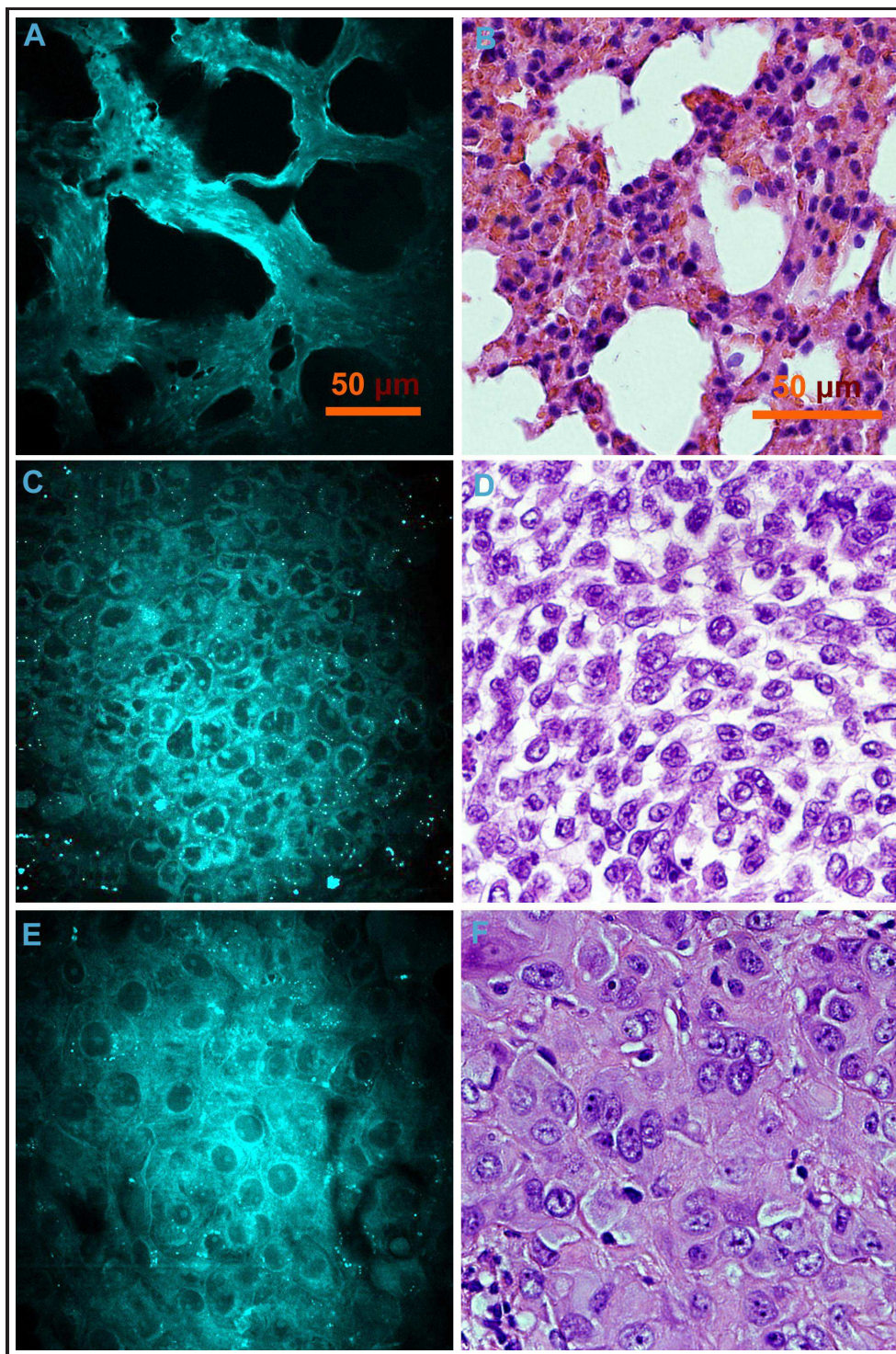


Figure 1.1 : CARS images and corresponding H & E images, of healthy tissue (A, B), adenocarcinoma(C,D), and squamous cell carcinoma (E,F)

In Chapters 2 and 3, we develop the image segmentation strategy into a fully automated cell nuclei segmentation framework from CARS images, and for three cell types of cancer: adenocarcinoma, squamous cell carcinoma and small cell carcinoma. Specifically we say cell nuclei segmentation and not cell segmentation, because the features needed to perform differential diagnosis reside with cell nuclei, and not the whole cell. This allows for the automatic and rapid information extraction from CARS images, making possible the development of an automatic diagnostic platform from such information.

In chapter 4, we propose a new method to extract and examine information from segmented CARS data to overcome the inability of previous data analysis methods in automatically differentiating NSCLC cell types. Figure 1.2 shows a general conceptual overview of a complete CARS based lung carcinoma diagnostic platform, the first part of which deals with the imaging physics, optics, and image acquisition, extensively developed in previous studies. Specifically, Our work is centered with the “Diagnosis” panel of the figure, where we aim to perform automatic segmentation of cell nuclei, and use the segmented information to perform separation of NSCLC cell types

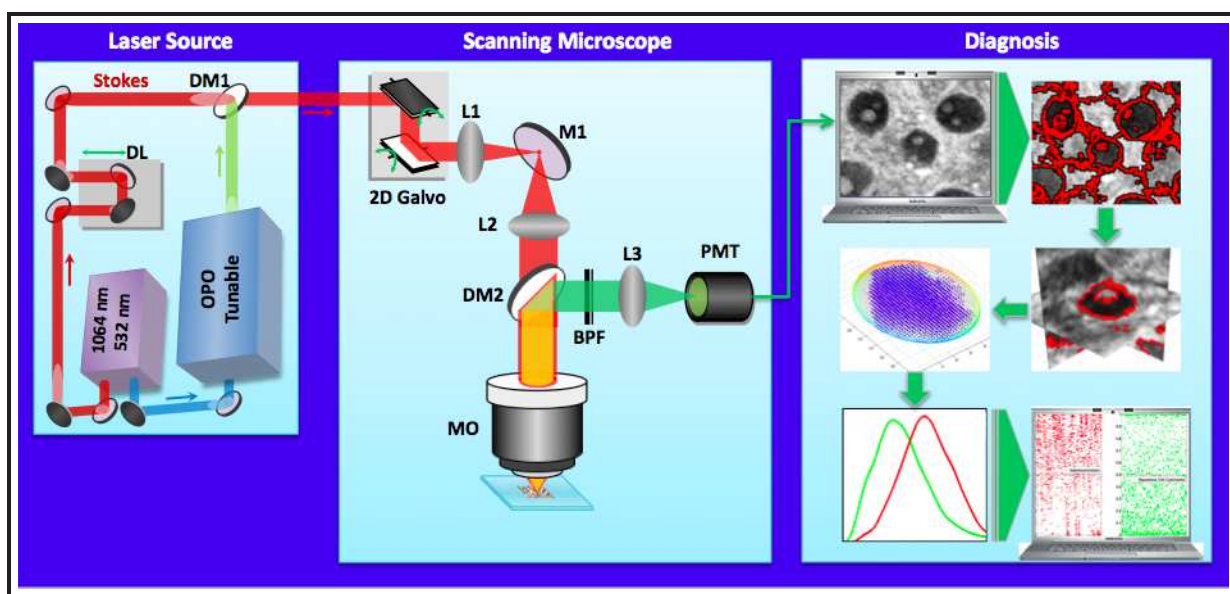


Figure 1.2 : Overview of a complete CARS based automatic lung carcinoma diagnostic platform

Chapter 2

Two Dimensional Automatic Segmentation of Cell nuclei

Previous studies attempting to perform automatic differential diagnosis of lung carcinoma relied heavily on segmentation processes that can be described as semi-automatic at best, requiring a human expert to select each cell nucleus to be studied [13,28]. Such approaches are far from optimal, and constitute a bottle neck to full automation. Automation is desirable from the standpoint that automatic location and delineation of cell boundaries can significantly save physicians and patients valuable time, while also reducing the cost of treatment and diagnosis. In addition, full realization of an on-the-spot endoscopic diagnostic system for lung cancer cannot be achieved without the ability to rapidly and automatically segment cell nuclei to prior to diagnosis.

This chapter is adapted from the publication:
Hammoudi, A. and Li, F. and Gao, L. and Wang, Z. and Thrall, M. and Massoud, Y. and Wong, S. T. “Automated nuclear segmentation of coherent anti-Stokes Raman scattering microscopy images by coupling superpixel context information with artificial neural networks”. 2011;317–325, Machine Learning in Medical Imaging [31].

2.1 The CARS Image Segmentation Problem

2.1.1 CARS Images and Standard Segmentation Approaches

CARS images are characterized by low signal-to-noise ratio (SNR), and uneven background which render most common and well performing segmentation approaches, that have been previously used to segment cellular nuclei, ineffective. For example, intensity thresholding, or adaptive thresholding, based on Otsu's method [32–34], are unable to separate those cells which are in contact, due to the low contrast in CARS images. Segmentation using marker controlled watersheds [5, 35, 36], graph cuts [6, 37, 38], or active contours [8, 39–41] all require initial seed points [5, 6, 8, 35–41] representing the centers of cells or nuclei. Such marker data is not readily available for CARS images and is usually determined either by a human expert, which violates the principal requirement of having a fully automated segmentation method. Another approach is to employ a tiered segmentation process that can detect the nuclei centers before performing delineation. Cell detection is yet another problem that we are trying to tackle as part of automating the segmentation process. Moreover, even with the availability of a solution to the cell detection problem the aforementioned edge delineation methods would still suffer shortcomings. Marker controlled watershed is susceptible to intensity variations, the low contrast and the uneven background possibly resulting in significant under-segmentation. The same problems are encountered by active contours and graph cuts, as CARS images quality will result in edge leaking and under-segmentation or early edge stop and under-segmentation.

2.1.2 Hierarchical Segmentation

It was demonstrated in [42], [43], and [44] that utilizing a tiered segmentation approach that starts with edge delineation rather than cell detection can significantly increase the overall cell segmentation accuracy in microscopic images. This is mainly merited to the ability to neglect the cell detection as a first step. Which in turn allows using data clustering algorithms such as k-means [45, 46], or expectation maximization [47], to exploit image features at the level of small neighborhoods, and create small image patches that over-segment the entire image while adhering to natural image boundaries, in the case of CARS images these include the boundaries of cell nuclei.

Following the creation of said image patches, the segmentation problem becomes a data classification problem requiring the ability to discriminate those patches that belong to objects of interest, in this case cell nuclei, from patches that correspond to the background or other objects, in the case of CARS images, these are the lipid tissue and reflective water molecules.

2.1.3 Context Based Hierarchical Segmentation

As such, we designed a hierarchical segmentation and detection approach that employs simple linear iterative clustering (SLIC) to partition CARS images into patches corresponding either to cell nuclei or background. Next, we designed a set of features to describe physical characteristics of the generated patches, and capture the image

context in which they exist. The idea of defining question behind the design of each of the descriptors, was “how would a person looking at the partitioned image tell which patches are nuclei and which are not?” And the answer is the context in which they reside. Those descriptors were used in training a classifier that can determine which patches belong to cell nuclei and which do not. The segmentation and detection framework is illustrated in figure 2.1.

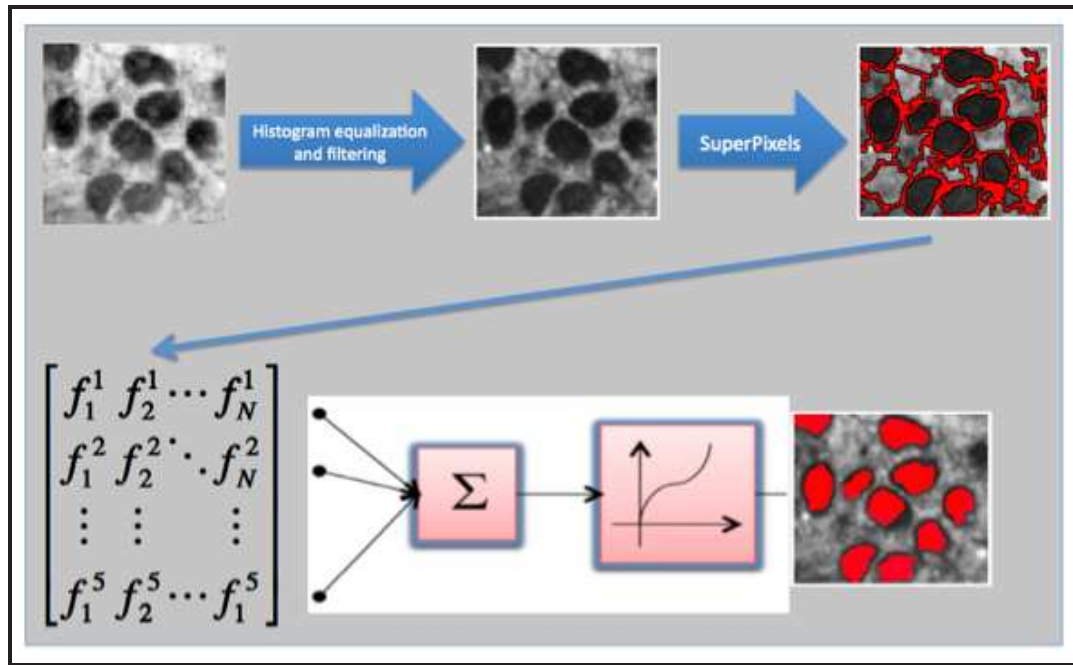


Figure 2.1 : The hierarchical approach used for segmentation and detection of cellular nuclei in CARS images

For the remainder of this chapter we present the details of the designed segmentation approach, local clustering is explained in section 2.2 and context based nuclei detection in section 2.3. A discussion of the results is provided in section 2.4.

2.2 Local Clustering in CARS Images

The use of image patches created using both local intensity and location information has been demonstrated to enhance the segmentation of color images, especially those with small closely positioned cells, low SNR, and uneven background [42–44, 48]. Superpixels, which are small image patches constituted of irregularly shaped connected groups of pixels, and generated by applying a clustering algorithm to localized regions of an image have been demonstrated to have superior performance in segmentation tasks to that of using rectangular image patches [43, 44, 48]. The advantage in using superpixels over rectangular patches is the ability to control the rigidness in their shape, and balance the dependence on position information to that of intensity information, this allows the superpixels to adhere well to the natural boundaries in images and capture their fine variations. The boundary of a superpixel containing a cell nucleus, for example, adheres very well to the boundary of that nucleus, figure 2.2. Moreover, the computational load of analysis tasks farther down the data analysis pipeline is reduced by the reduction in the number of data points, as any further analyses will be performed on clusters of pixels, treated as a single data point, rather than individual pixels. For the remainder of this discussion the terms superpixel and cluster convey the same meaning and will be used interchangeably.

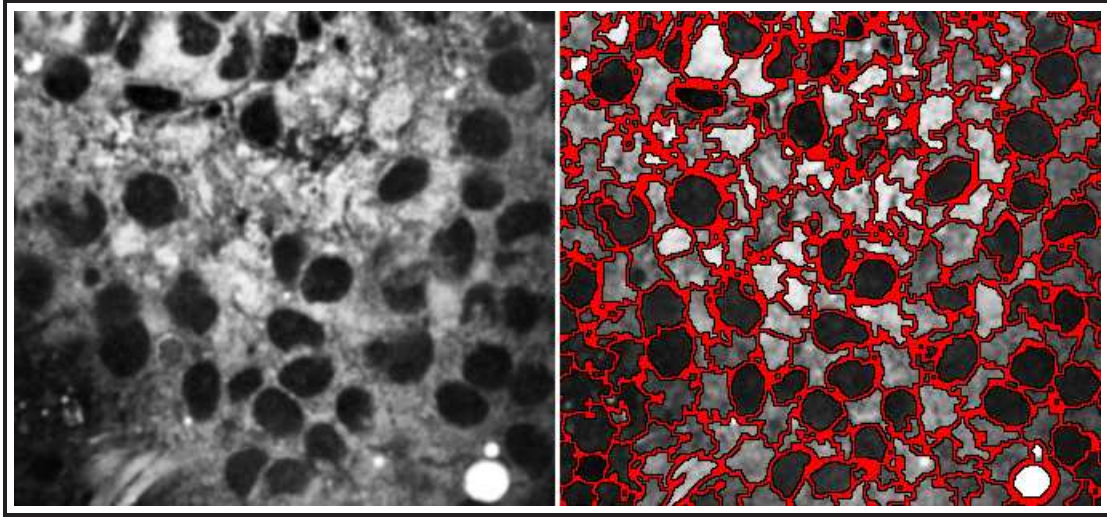


Figure 2.2 : Superpixels generated in a CARS image of an Adenocarcinoma sample. Red markers represent the boundaries of superpixels. It is notable how well the superpixels adhere to image objects, namely nuclei

2.2.1 A Primer to Simple Linear Iterative Clustering Based Superpixels

Multiple algorithms for creating superpixels exist, such as [44,48,49], but the simple linear iterative clustering (SLIC) based algorithm has shown superior performance in terms of little under-segmentation and over-segmentation errors, image boundary detection, and computational complexity [48]. Specifically, it features superior boundary detection capability, along with a low computational resource requirements, both of which fit well for CARS image segmentation, where accurate nuclear edge detection and use of a resource-friendly algorithm for rapid processing of large data sets are equally desirable.

SLIC operates as a modified k-means algorithm that performs clustering on a subset of data points confined to small regions of interest in a data set, rather than

partitioning data sets into k clusters to which any data point can belong, like k -means does. Thus, a pixel, which is represented by a feature vector whose components are the coordinates and color information of that pixel, does not potentially belong to any of the k clusters, but instead to a small subset of clusters in the neighborhood of that data point. Using this strategy, pixels having similar characteristics are grouped together, but the limited search area ensures that pixels that are far apart, such as very similar pixels belonging to two different nuclei, are never in the same cluster. This reduces the possibility of multiple nuclei belonging to the same cluster and, at the same time, supports for the assumption that later processing steps can treat a cluster as a single nucleus.

To perform SLIC clustering and generate superpixels, every pixel p in a volume is represented by a vector of features $f_p = [L_p, A_p, B_p, x_p, y_p]^T$ where L_p , A_p , and B_p represent CIELAB color space values at p . x_p , and y_p , represent the coordinates of p in a CARS image. We present a summary the SLIC superpixel clustering that we performed in Algorithm 1 adapted from [48]. In Chapter 3 we show our extension of SLIC to perform clustering in higher dimensional data sets.

Algorithm 1 SLIC Superpixel creation

Set initial k cluster centers by sampling the image with a uniformly spaced grid:
 $C_k = [L_k, A_k, B_k, x_k, y_k]^T$, with grid separation S

repeat

for each pixel p **do**

 Compute distances to each neighboring cluster center $C_n, n \in [1, 2, \dots, 8]$

end for

 Assign each pixel to the cluster with whose center is the nearest.

 Compute new cluster center positions

 Compute E

until $x < threshold$

Where the algorithm parameters are defined as follows:

- k is the number of clusters, or superpixels, to be generated, it was determined empirically, to be 400 for a 512×512 CARS image. This was determined through an exploration of the parameter design space, detailed in subsection 2.2.2.
- The separation S between two grid points, determined both by the size of the image, and the number of super pixels and defined as: w/\sqrt{k} , where w is the width of the CARS image.
- E is the residual error after every iteration computed as:

$$E = \sum_{j=1}^k \|\mathbf{C}_{j1} - \mathbf{C}_{j0}\|_{\ell_1} \quad (2.1)$$

Where \mathbf{C}_{i1} is the vector representing the new cluster center computed at the end of the current iteration, \mathbf{C}_{i0} is the same cluster center from the previous iteration, ℓ_1 is the $L1$ vector norm, and k is as defined above. The threshold

was Set to 0.5, details of determining the threshold are presented in subsection 2.2.2.

- Distance between 2 pixels in the 5D feature space used for segmentation is computed as follows:

$$d = \sqrt{\frac{(L_{p1} - L_{p2})^2 + (A_{p1} - A_{p2})^2 + (B_{p1} - B_{p2})^2}{m} + (x_{p1} - x_{p2})^2 + (y_{p1} - y_{p2})^2} \quad (2.2)$$

$p1$ and $p2$ represent 2 pixels, and m is a parameter that controls the rigidity of the superpixel boundaries, large values for m produce superpixels that are more dependent on location and more square in shape, while smaller values, result in superpixels with more fluid boundaries. For 2D segmentation the value of m was set at 10, a value used in the literature. In chapter 3, we demonstrate how the value was determined for 3D segmentation.

- Finally, it is worth noting that the distance is computed from each data point to only cluster centers as part of the localized clustering approach, it is assumed that a pixel cannot belong to a cluster whose center is at a distance greater than S to that pixel. Hence, the only eligible cluster for a pixel to belong to are at most those having the 8 immediate neighboring cluster centers. This is illustrated in figure 2.3. The red circles represent pixels of interest, the red squares represent the boundaries of the search area defined by S , and the blue crosses represent the cluster centers within the search area. (A) point is positioned at

a distance S from edge of search region, 8 clusters are eligible. in (B) the same point undergoing only a horizontal, or vertical translation, potentially belonging to 6 clusters, and undergoing both a horizontal and vertical translation in (C) the point potentially belongs to 4 clusters.

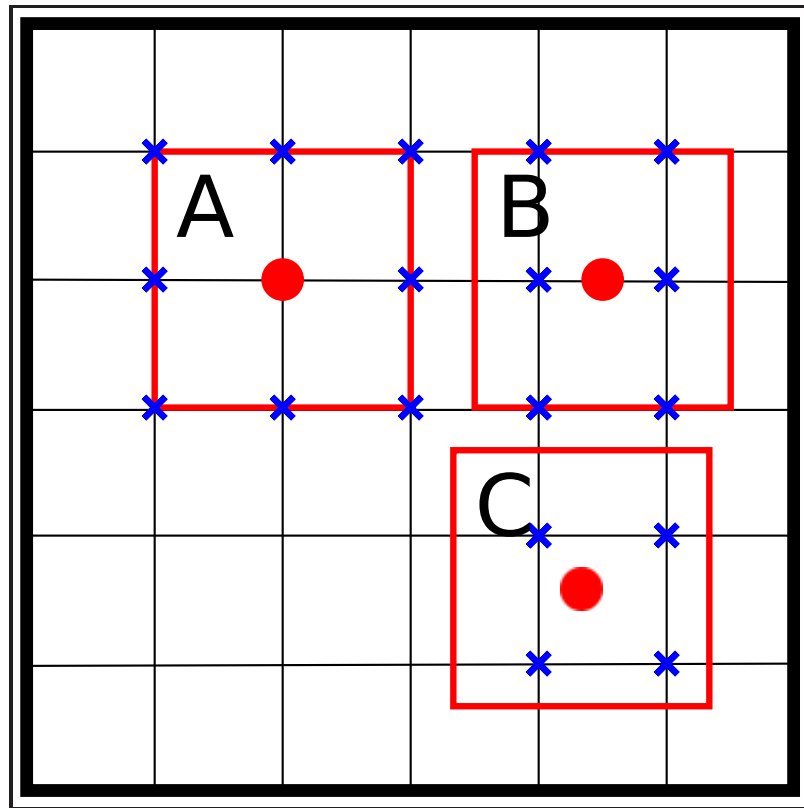


Figure 2.3 : Illustration of potential data point-to-cluster assignment.

2.2.2 Experimental Considerations of Superpixel Segmentation in CARS images of Various Lung Carcinoma Subtypes

As described, the SLIC algorithm requires two user inputs, namely the number of superpixels or clusters k , and the fluidness parameter m . Because there is no formal

method to determine those parameters, we ran a rough exploration of the parameter space, performing the clustering process on a test CARS image for multiple values of k and m , as show in figure 2.4. We observed the following:

- We observed that a large numbers of clusters, with a small value of m result in the best clustering results in terms of nucleus boundary detection. However generating a large number of clusters is resource intensive, and results in the unfavorable over-segmentation of nuclei. The latter effect would add the tedious requirement of detecting small clusters as parts of nuclei and merging those clusters together in the subsequent detection stage.
- A small k , and a large m will result respectively in under-segmentation where clusters will include including more than one nucleus or nuclei and background. and rigid cluster boundaries, that are more dependent on the spatial components of the feature vectors, and subsequently rectangular in shape, with little or no representation of the actual image edges.
- Very small values of m will result in extremely fluid boundaries that are sensitive to small intensity variations caused by noise in the images, and considerably diverge from natural image boundaries.

Subsequently, we opted for a value of k that would produce just enough clusters such that a cluster is large enough to contain a complete nucleus and nothing else. i.e. every nucleus is contained in a single cluster, and the cluster itself is confined to

that nucleus. that value of k was 300 clusters. The same study was performed for 3D CARS images, chapter 3, 2000 clusters were used for those data sets. In a similar manner, the value of m was set at 10.

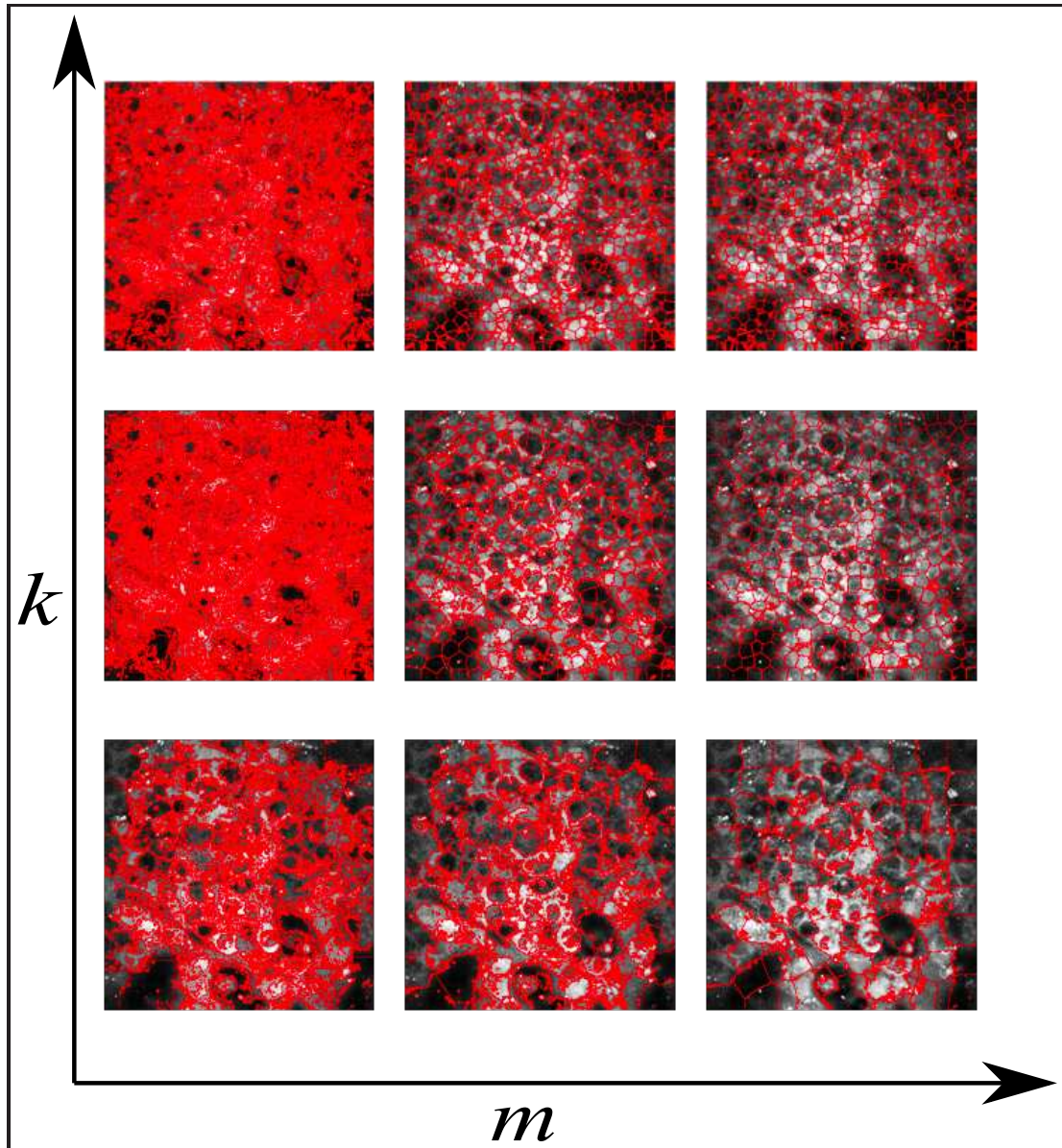


Figure 2.4 : Design space exploration: effect of varying m and k on the generated superpixels

Examining the change in the residual error over every iteration, it was noted first, that the residual undergoes exponential decay decreasing significantly in the first 5 to 10 iterations, then decreasing rather slowly after that, with little significant decrease beyond the value of 0.5, figure 2.5. Henceforth the threshold for E beyond which the algorithm terminates, was set at 0.5.

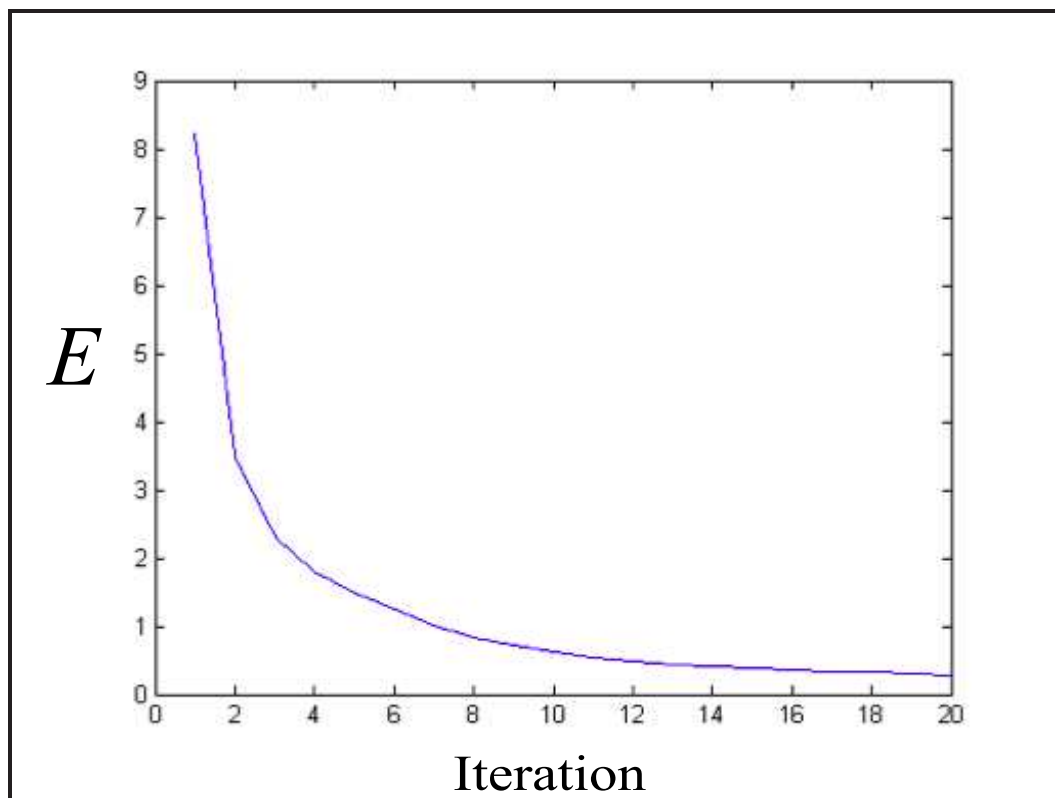


Figure 2.5 : Convergence of SLIC clustering in a CARS image

Figures 2.2, 2.6, and 2.7 show CARS images of the 3 carcinoma cell types of interest, adenocarcinoma, small cell carcinoma, and squamous cell carcinoma, respectively. The left panels represent the original images, and the right panels, show the images with the cluster boundaries marked in red after generating superpixels.

In all three cell types it is notable that superpixels clustering delineates nuclear boundaries, and generates clusters that contain only either cell nuclei or background, effectively solving the edge delineation problem and reducing the CARS image segmentation problem into one of determining which clusters correspond to cell nuclei, and which correspond to image background or other structures.

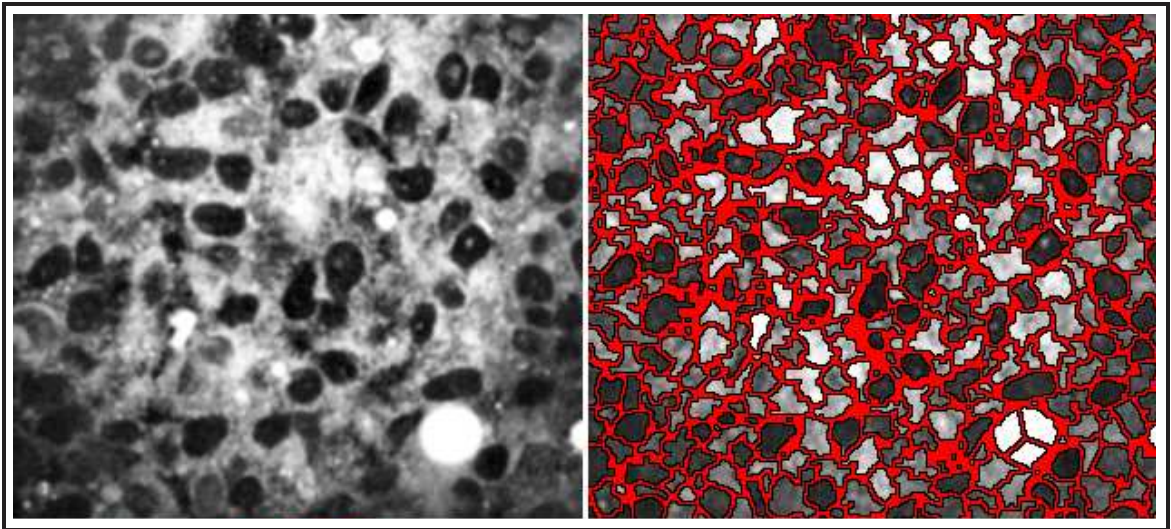


Figure 2.6 : Superpixels generated in a CARS image of a Small cell carcinoma sample. Red markers represent the boundaries of superpixels.

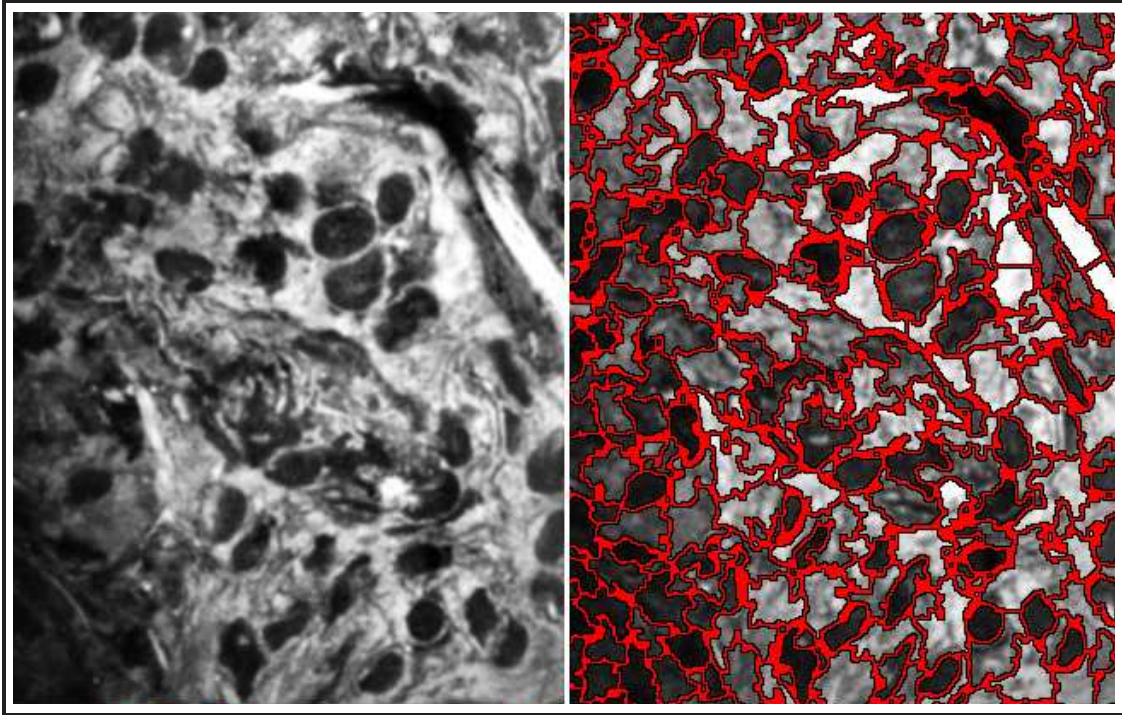


Figure 2.7 : Superpixels generated in a CARS image of an Squamous cell carcinoma sample. Red markers represent the boundaries of superpixels

Before concluding the discussion on clustering, it is worth noting that all images were intensity adjusted using adaptive histogram equalization [50,51] de-noised using a 7×7 Gaussian smoothing kernel [52,53] of the form

$$G_{\sigma}(x, y) = \frac{1}{2\pi\sigma} \exp \left\{ \frac{-(x^2 + y^2)}{2\sigma^2} \right\} \quad (2.3)$$

and illustrated in figure 2.8

A final note is that all the CARS images used for 2D segmentation are 512×512 images. In addition, for this, and all subsequent CARS image related computations, in 2D or 3D, we use the physical dimensions of the pixels to compute coordinates.

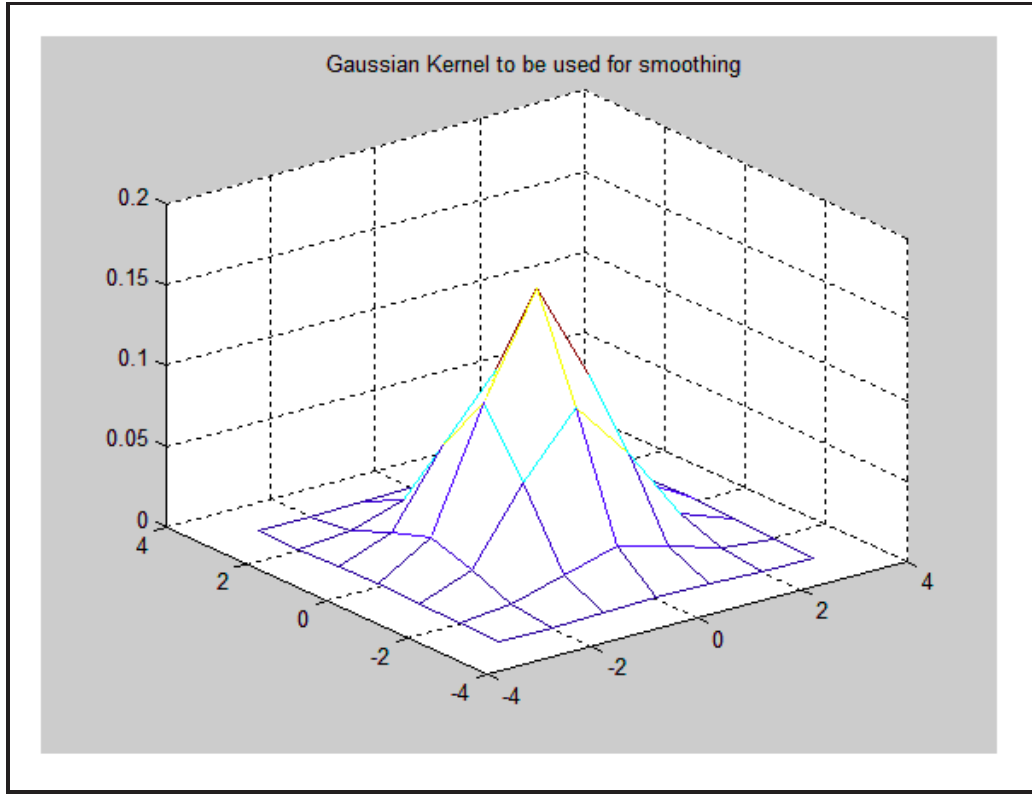


Figure 2.8 : Gaussian filter used to de-noise CARS images

The physical dimensions are simply determined by the instrumentation setup, they are square pixels with sides of length $0.306904 \mu m$, and the transformations are as follows:

Pixel indices to physical location transformations.

$$x_{phys} = 0.306904x_{pi} \quad (2.4a)$$

$$y_{phys} = 0.306904y_{pi} \quad (2.4b)$$

In section 2.3 we describe the features we designed to measure the contextual information of the superpixels, and used to train a classifier to determine which clusters are cell nuclei and which are not.

2.3 Using Superpixel Context Information to Detect Nuclei in CARS Images

The second and final step to solving the nuclear detection and segmentation problem, is building a classifier capable of discriminating those superpixels that contain nuclei from those that contain other image objects. We proceed to this step under the assumption that the clustering algorithm performs well in generating superpixels that adhere well to natural image boundaries, figures, 2.2, 2.6, and 2.7. As described in subsection 2.1.3, we aim to design features that reflect the human perception process of the nuclei in the images, and thus depend on the context in which the cells are perceived through, and the uniformity in the shape of superpixels corresponding to cells as opposed to the irregular shape of those belonging to the background. Those features were used to train an artificial neural network (ANN) classifier that could detect nuclei with great accuracy, the details of the features and the classification process are below.

2.3.1 Defining Context Descriptors

It is difficult to discriminate nuclei from background only using the intensity information of individual superpixels. Since the nuclear superpixels have, in general, regular shapes, and low and uniform intensity distribution compared to the background superpixels, the nuclei within a small neighborhood are distinguishable mainly through two major factors: The intensity variation of a superpixel with respect to its neighbors, and the shape uniformity. Thus we define the superpixel context index for distinguishing the nuclei from background.

We computed two superpixel context indices as follows. For each superpixel, locate all of its immediate neighbor superpixels, and then calculate the ratios of the average and median intensity of the superpixel to the average and median intensities of its neighboring superpixels respectively. We set the superpixel context indices as the minimum ratio values (mean and median) based on the fact that at least one immediate neighbor is a background superpixel. A summary is given in algorithm 2 below

Algorithm 2 Computing superpixel context indices

```

for each superpixel  $P$  do
  Locate all neighbor superpixels  $P_{nj}$ 
  compute:  $\min \left\{ \frac{\mu_{IP}}{\mu_{IP_{nj}}} \right\} \forall_j$ 
  compute:  $median \left\{ \frac{\mu_{IP}}{\mu_{IP_{nj}}} \right\} \forall_j$ 
end for

```

To measure the shape uniformity, we designed a descriptor based on ray bursts [54],

as follows: locate the centroid of a given superpixel, shed rays to the boundary points uniformly (each at a 10° arc), and use the standard deviation of lengths of the 36 ray lines as the uniformity value.

In addition, we added two other relevant descriptors that can demonstrate shape uniformity. The first one is the goodness of fit between the superpixel and the best fitting ellipse. We fit each superpixel with an ellipse, and then calculate the relative regions outside the ellipse and missing regions inside the ellipse compared to the ellipse size. The second is the ratio of superpixel area to the area of its convex hull, equation 2.5. Figure 2.9 illustrates ray descriptors and the goodness of fit descriptor.

$$\frac{Area_P}{Area_{ConvexHullP}} \quad (2.5)$$

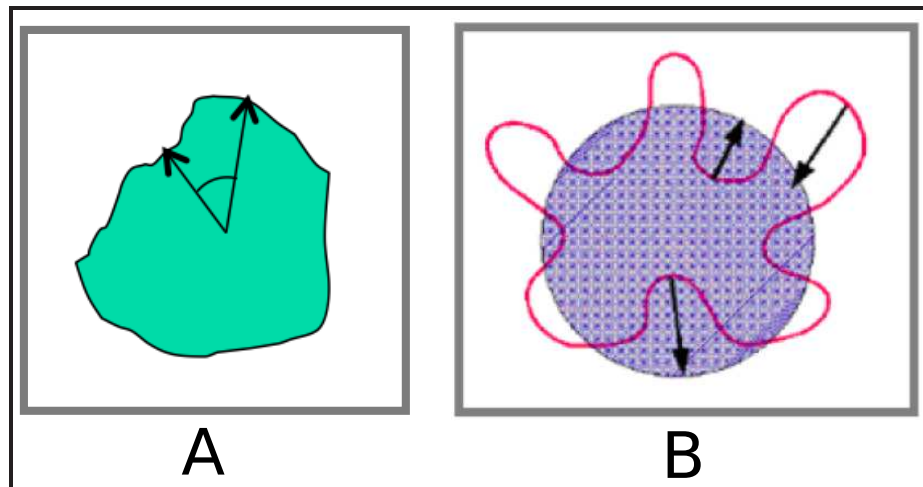


Figure 2.9 : Illustration of ray descriptors (A) and Ellipsoidal goodness of fit(B)

2.3.2 Nuclear Superpixel Identification with an ANN

The selected features are used to create a feature space to be used in classification, it can be described by the vector $[f_1, f_2, \dots, f_5]^T$ whose components are the morphological features detailed above. We use a feed-forward back propagation neural network [52, 55], which could learn the optimal feature combinations for the nuclear superpixel identification. Artificial neural networks are an extensively researched topic of which we provide a summary of the basic operating principals herein, and provide specific details necessary for the replication of this work.

A neural network is a function fitting or pattern recognition paradigm loosely inspired by the human learning process, it can learn to produce the correct outputs for a given set of inputs or stimuli, by adjusting the weights of the connections between its building blocks or neurons [52, 55]. As the naming suggests, a neural network is an interconnected set of neurons, a neuron is depicted in figure 2.10 below.

The neuron depicted in figure 2.10 receives a vector input of real numbers $[i_1 \dots i_5]^t$. Note that a neuron can have less than or much more than 5 inputs, we use 5 here for illustrative purposes because it is the size of our superpixel feature vectors. Each input is weighted with another real number, weights here are represented by the vector $[w_1 \dots w_5]^t$. the output n of the summing block, a linear combination of the inputs plus the bias b , is processed by the output transfer function f to produce the final output o . f serves the purpose of changing the output from an arbitrary real number to one of 2 possible output classes, such as 0 or 1. The operation of a neuron can

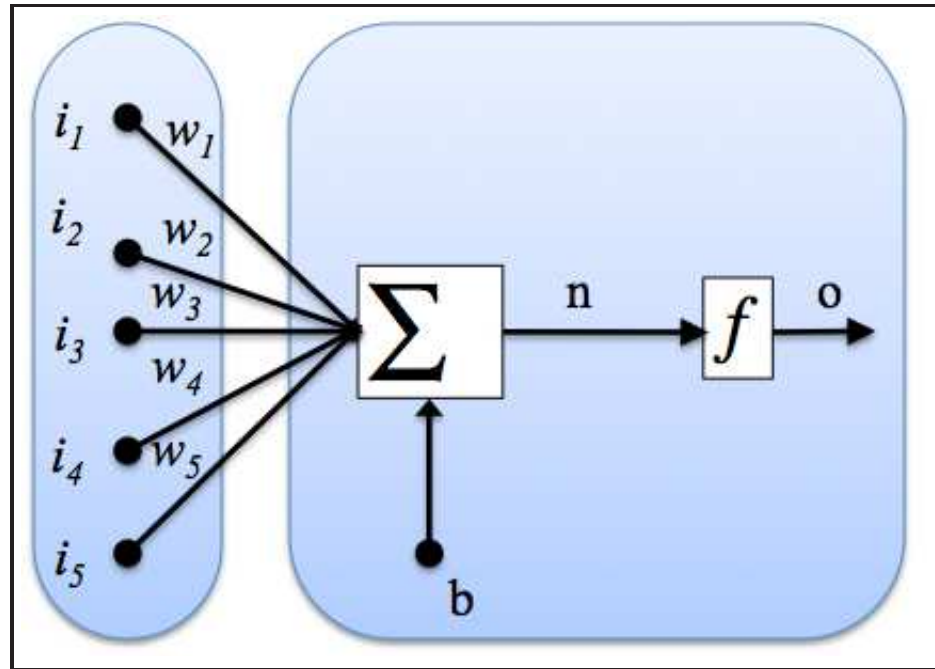


Figure 2.10 : Schematic representation of a neuron

be described by 2.6. The neuron is taught, or *trained* to predict the correct output to given inputs using a set of pre-collected and labeled data, which is comprised of a set of input vectors and their known outputs. For example, in the case of nuclear superpixel identification, training is done using a set of feature vectors of superpixels, and their corresponding labels, done manually, as background or nuclei.

$$o = f(\mathbf{wi} + b) \quad (2.6)$$

During training, the neuron is fed with the training inputs, and the output is compared to the correct expected target - training - output, following which, the weights and the bias are adjusted to minimize a measure of error between the actual

and desired outputs bringing the two as close as possible. this is repeated for all training inputs. When the neuron is presented with an a new unknown input, not belonging to the training set, it's output is expect to be a correct classification. This learning process can be made more powerful, and the accuracy in classification can be arbitrarily increased through cascading and layering neurons [55,56]. A schematic depiction of a multilayer neural network is depicted in figure 2.11, in which one layer comprised of multiple neurons feeds outputs from those neurons to the next layer in the network.

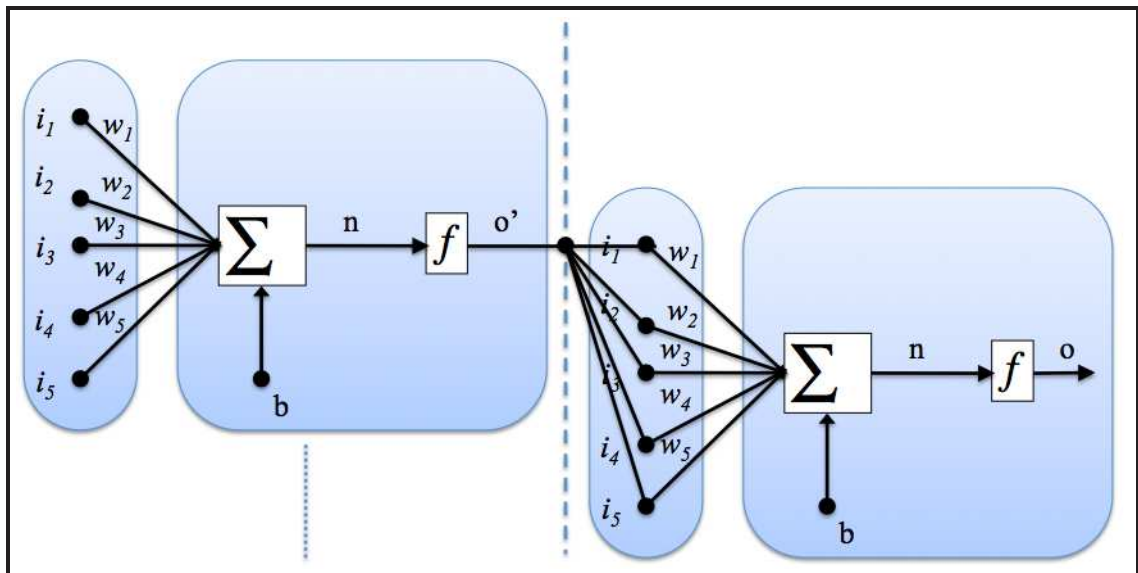


Figure 2.11 : Schematic representation of interconnected neurons in a multilayer neural network

Training a multilayered network is done in the same fashion as a single layer network, in which inputs are fed to the first layer and the output of the last layer is compared to desired target outputs, followed by adjusting the neuron weights and

biases to achieve minimal error between actual and desired outputs.

In this study, a neural network with 100 neurons and one hidden layer is trained for each of the lung cancer subtypes. An output layer is added to convert the outputs from all the neurons in the hidden layer into a single output, in this case the classification result, a 1 or a 0. Training was done with the back propagation algorithm that minimizes squared error between network outputs and target outputs. [55, 56]. The output function of the neurons was the sigmoid function [56] of the form 2.7 which insures the output has a value between 0 and 1.

$$o = \text{logsig}(n) = \frac{1}{1 + e^{-n}} \quad (2.7)$$

The neural network was built using the Mathworks' MATLAB[®] neural network toolbox. The feature vectors were computed for every superpixel in a CARS image, and combined as shown in the lower panel of figure 2.1, the classifier uses the features to classify superpixels as nuclei or non -nuclei superpixels. To train the classifier, we manually labeled the superpixels (as nucleus or background) in three images for each subtype of lung cancer to train the three ANN classifiers respectively.

Figures 2.12, 2.13, and 2.14, show segmentation and detection examples in images of each of the three lung carcinoma cell types. Each of the figures shows the original images and superpixels in the upper panels, and the lower panels depict the superpixels that were classified as nuclei, and shows their boundaries. It is notable how well the approach detects and delineates the nuclei. We tested the approach on

a random set of previously labeled CARS images, and computed validation scores, precision, recall, and f-score, which are presented in table 2.1.

2.3.3 Results

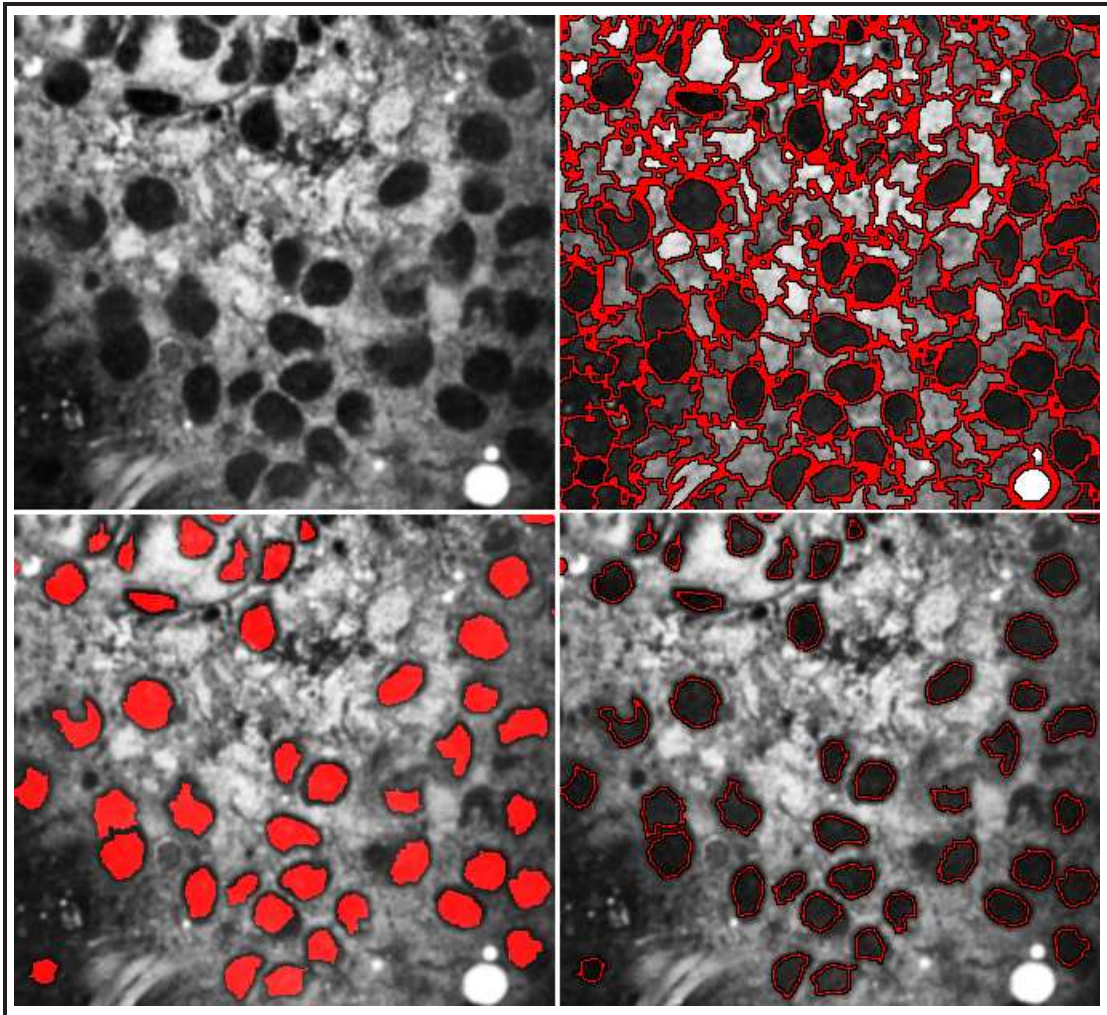


Figure 2.12 : Segmentation and detection example: adenocarcinoma sample

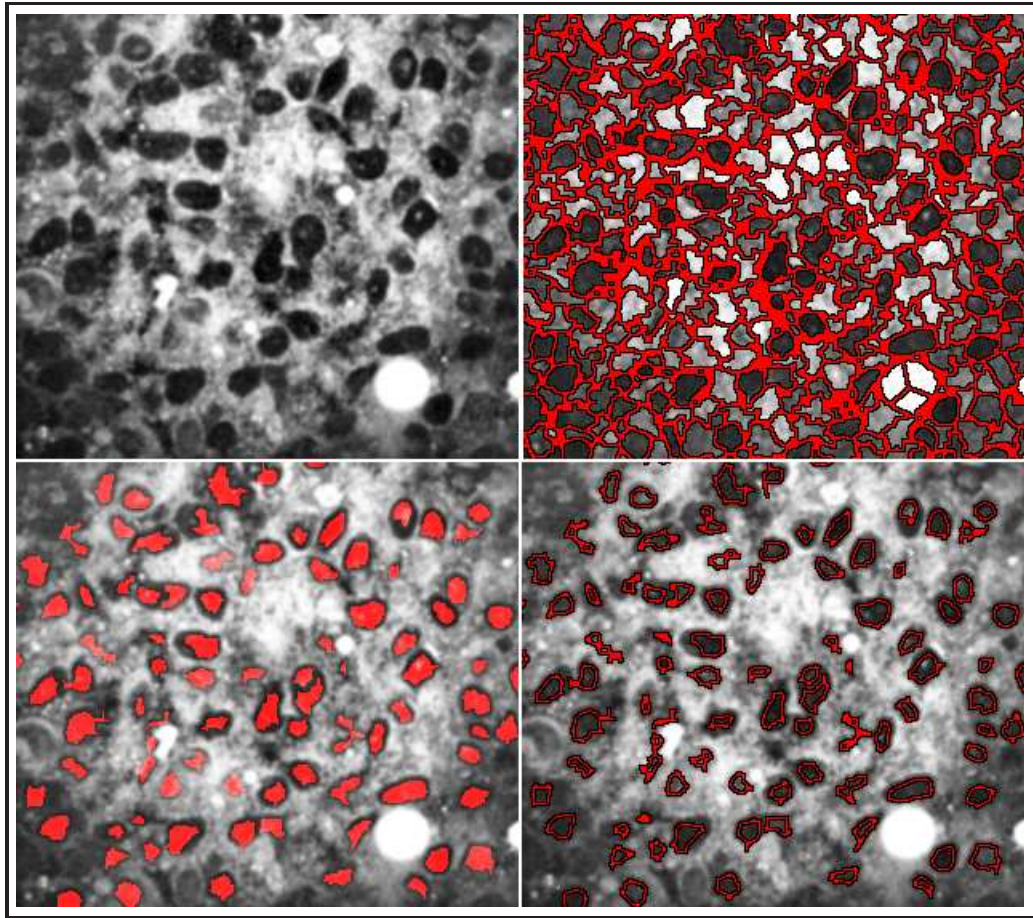


Figure 2.13 : Segmentation and detection example: small cell carcinoma sample

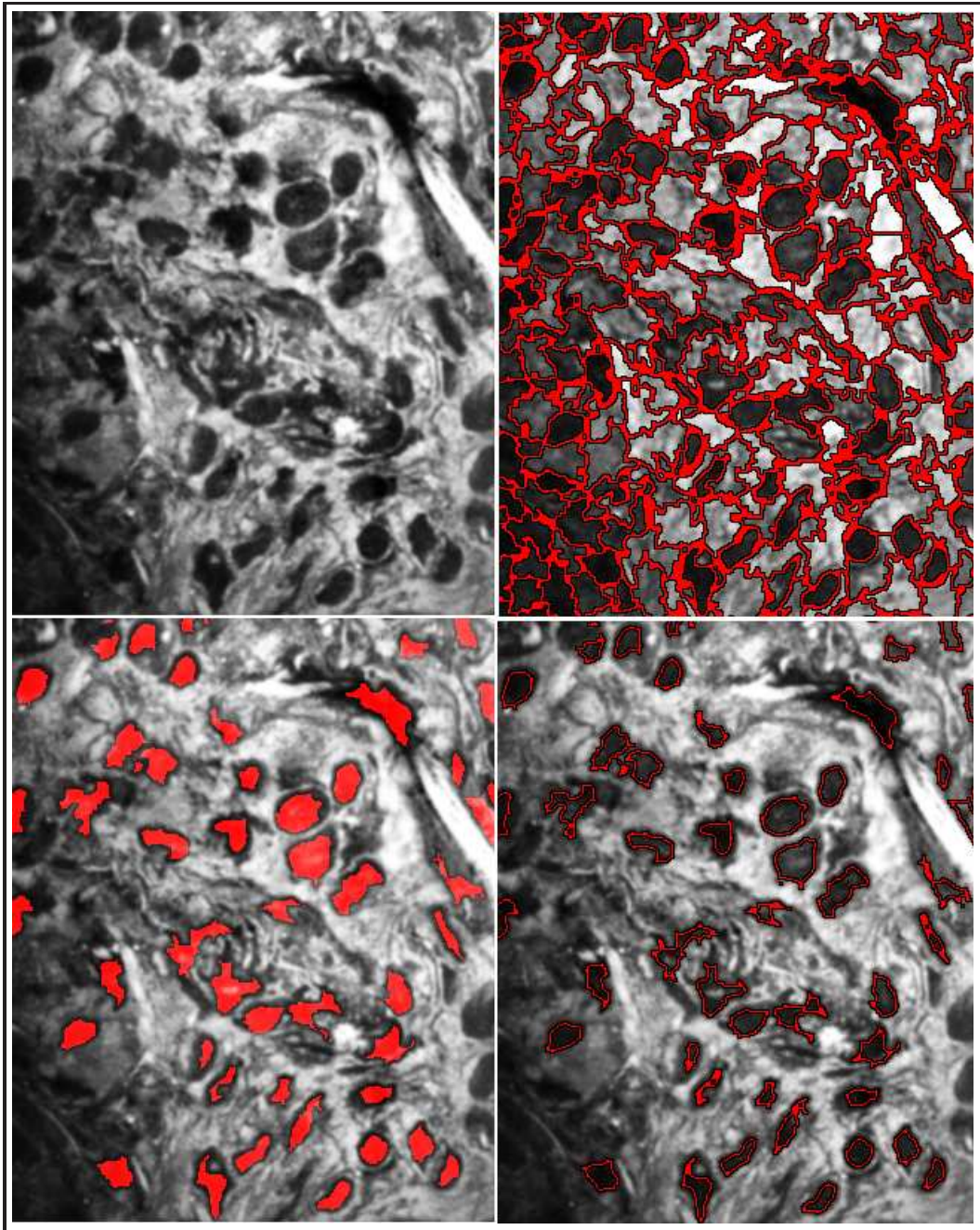


Figure 2.14 : Segmentation and detection example: squamous carcinoma sample

Table 2.1 : Validation scores for the segmentation and detection framework in all lung carcinoma subtypes

Cell Type	Precision	Recall	f-Score
Adenocarcinoma	94.8%	93%	93.8%
Small cell carcinoma	92.1%	92%	92%
Squamous cell carcinoma	93%	95.6%	94.8%

2.4 Discussion

Thus far, we proposed a fully automated 2D segmentation method for cellular nuclei in CARS lung carcinoma images, coupling superpixels (local clustering) and ANNs. Thus, we provided a solution to a critical step for differential analysis using cell morphology with accuracy that is $> 90\%$. The experimental results demonstrated that the proposed automated method possesses excellent capability for nuclear segmentation on CARS images. To get to the full potential of using CARS images for differential analysis of lung carcinoma cell types, the local clustering and classification technique was extended to 3D CARS images, where biologically significant features can be measured with more fidelity than 2D images to classify segmented cells into their respective lung carcinoma cell types. This is detailed in chapter 3.

Chapter 3

Three Dimensional Automatic Segmentation of Cell nuclei

3.1 3D Segmentation: Why? And How?

3.1.1 introduction

Thus far, we have established a solution that addresses the first issue raised in the introduction, namely that previous work has relied heavily on human intervention in order to segment 2D CARS images for use in differential analysis. However, the end goal of our study is not the segmentation itself, rather, it is the use of segmented images in order to extract pathologically relevant features that would solve the differential diagnosis problem of non small cell lung carcinoma cell types.

Moreover, although previous studies did not in fact develop a fully automated method, they did have access to a set of manually segmented CARS images large enough to be used in tackling the diagnosis problem [13,28]. Nonetheless, diagnosing NSCLC using data extracted from 2D images proved to be too difficult, because 2D

This chapter is adapted from the technical report:
Ahmad Hammoudi, Yanqiao Zhu, Fuhai Li, Liang Gao, Michael J. Thrall, Jinwen Ma, Yehia Masoud, Ming Zhan, Zhiyong Wang, Stephen T.C. Wong, “Automated 3D Segmentation of Cell Nuclei in Coherent Anti-Stokes Raman Scattering (CARS) Microscopy Images to Enable Label Free Cancer Subtyping”. Technical Report, The Methodist Hospital Research Institute, Department of Systems Medicine and Bioengineering.

slices of images of adenocarcinoma and squamous cell carcinoma look very similar, and the features used for diagnosis, are measured in a setting that is a distorted representation of the environment in which cells reside in inside a tissue. For example, measuring the size of a nucleus in a 2D image is done through measuring its area, however, a nucleus is a three dimensional object whose size is represented by volume. Hence measuring size in a 2D image reduces to measuring the area a nucleus occupies in a given slice of the image stack, which can vary from a single point, to the entire circumference of that nucleus depending on where the slice was taken. Another major distortion that is introduced from 2D based measurements, takes place when measuring distances between nuclei. In the 3D setting distance between 2 nuclei is assumed to be the distance separating their centers, in a 2D measurement however, every slice on a nucleus is treated as a separate nucleus, and the distances measured in image slices, can possibly carry no representation of the actual distances in the tissue. In addition, when computing features in a large set of 2D images, the measurement will include a lot of redundancies, as each nucleus traverses multiple slices and will hence be included several times, very few of which carry significance to the real physical setting of the tissue, for the two reasons stated above. These distortions render NSCLC classification from 2D data a cumbersome problem to solve.

With this in mind, our solution proposes computing the nuclei's pathologically relevant morphological features in the cells' natural environment, i.e. in 3D images of the entire tissue, as opposed to studying individual slices. Such a measurement takes

full advantage of the ability of the CARS microscope to perform tissue sectioning. Through stacking all the individual slices from a CARS experiment and interpolating in between consecutive slices we can build a 3D image of the cells and the cancerous tissue, we call this a “volume”, in which we can perform real, undistorted measurements of the morphological features of interest. Figures 3.1, 3.2, and 3.3, respectively show a maximum intensity projection rendering of a CARS volume, a cross sectional image from all 3 directions of the same volume, and a close up to the volume in which the nuclei are visible in cross sections across all dimensions.

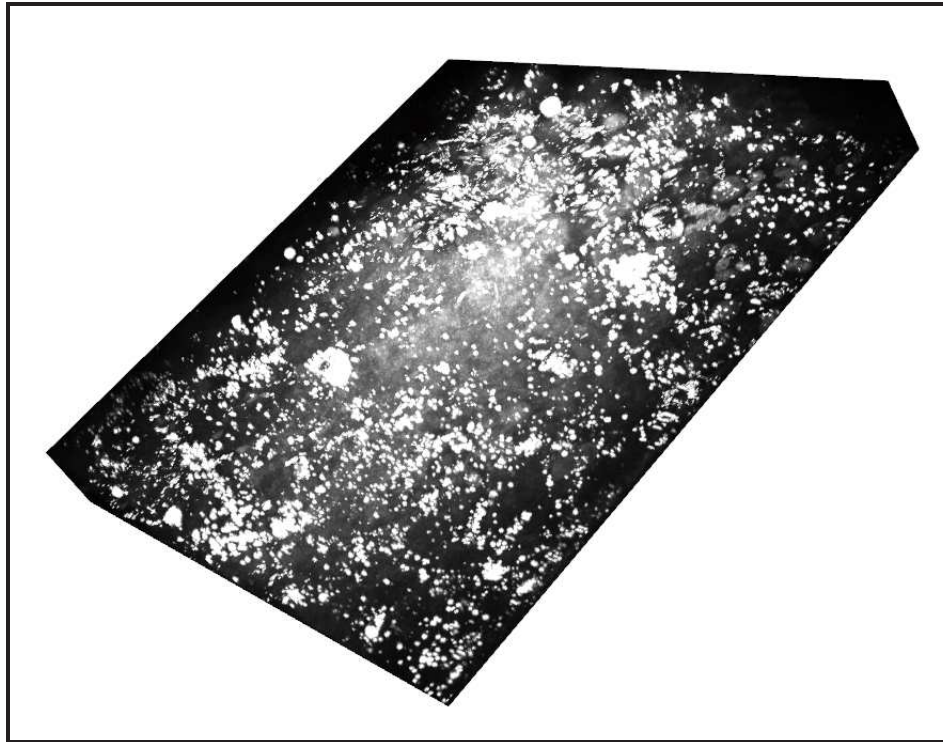


Figure 3.1 : A 3D CARS image stack of a lung tissue containing adenocarcinoma cells in its raw form

To extract pathologically relevant features from the nuclei, we needed segmented

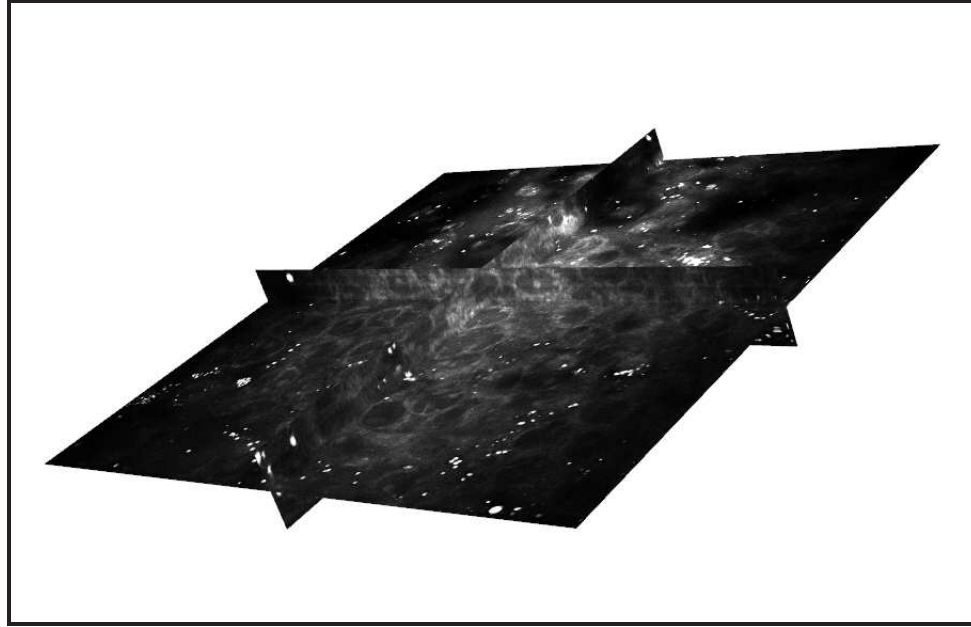


Figure 3.2 : A 3D cross sectional representation of a CARS volume

3D CARS, thus in this chapter we explain the details of extending the hierarchical automated nuclei segmentation and detection system, detailed in chapter 2 to operate on 3D CARS volumes similar to those shown in figures 3.1, 3.2, and 3.3. The end results of this are 3D volumes containing information about the exact shapes and locations of the cellular nuclei in the studied tissue, as measured by the CARS microscope. This chapter is organized as follows, the rest of this section discusses the 3D segmentation problem in general and describes our solution. In section 3.2 we show how the superpixel clustering process can be extended to any dimension, and we demonstrate it's use for segmenting 3D CARS volumes. Section 3.3 presents the extension of the features used in identifying nuclei to three dimensions, and details the classifier used to perform nuclear detection. The chapter concludes with a

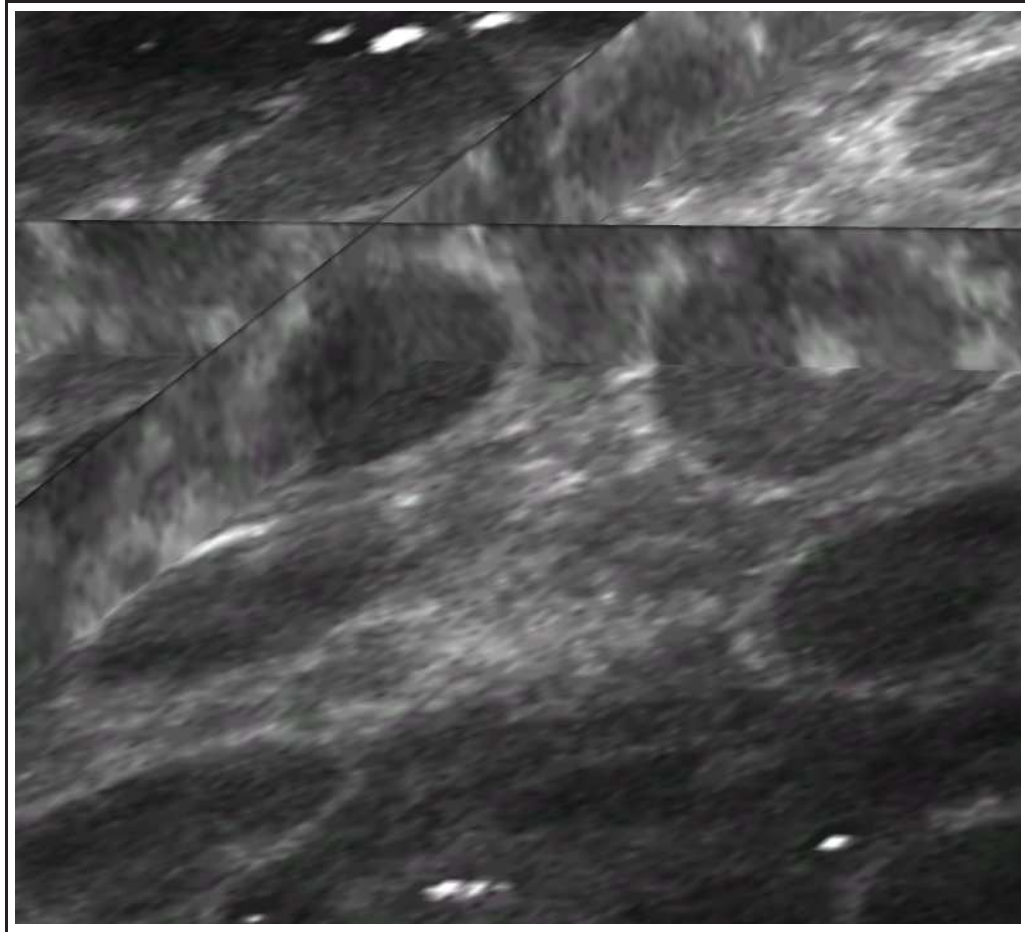


Figure 3.3 : A close up of a 3D cross sectional representation of a CARS volume, with nuclei visible in all 3 dimensions

presentation of the segmentation results and a brief discussion in section 3.4

3.1.2 The 3D CARS Image Segmentation Problem

The automated 3D detection and segmentation of cellular nuclei in CARS images remains an open problem suffering from the same impediments as those in 2D image segmentation such as low SNR, and noisy and uneven background. In spite of the fact that an array of segmentation approaches is available, none of those can effec-

tively perform segmentation of nuclei in CARS volumes without heavy modification or significant human intervention. We go again over some examples, thresholding based methods [32, 57] are not suitable due to the noisy and uneven background. Simple watersheds [58] will cause many false positives in nuclei detection. Adaptive Voronoi and seeded water-sheds could reduce the false positives by adding seed markers [59, 60], but no marker information is readily available in CARS images. Active contours, level sets, and graph cut methods [37, 39–41, 61, 62] could determine accurate object boundaries, but they require initial boundaries or seed markers, both of which are challenging information to acquire, that we address in the detection part of our hierarchical system .

In this study, we propose the first method of fully automated 3D cell nuclei segmentation of CARS images for enabling the subsequent differential diagnosis of NSCLC cell types. Extrapolating from the 2D segmentation scenario, first, we applied local clustering to partition 3D volumes into blocks that contain either nuclear regions or background in order to overcome the low SNR and uneven background. Second, we employed semi-supervised machine learning [63] to identify those blocks corresponding to cell nuclei based on a set of specifically designed morphological features. A framework resembling the 2D case framework shown in figure 2.1. In addition, because the end results of 3D segmentation are critical to the final differential diagnosis end goal, we demonstrate the effectiveness of our 3D segmentation approach by benchmarking its results against other well established segmentation and detection methods.

3.2 2D to 3D: Superpixels to Supervoxels

It was demonstrated in [64] and [65] that superpixels can be generated in any dimensions, by a simple extrapolation of the clustering algorithm used to generate them in two dimensions, to accommodate the additional spatial features of feature vectors in higher dimensions. In the case of 3D data sets, the clustering algorithm needs to accommodate an additional component, representing the 3rd spatial dimension, in the feature vector. In addition, when searching for candidate clusters to assign data points, or voxels to, the number of neighbors increases from 8 in the 2D case, to 26 in the 3D case. In 3D images, the generated clusters are dubbed supervoxels, as opposed to superpixels in 2D.

Figures 3.1, 3.2, and 3.3 above show that the images of interest are noisy and characterized by low contrast around the edges of the nuclei, the nuclei are even somewhat indistinguishable to the human eye. To circumvent these problems we used local clustering, i.e., supervoxel analysis, which performs clustering locally using both intensity similarity and location constraints, limiting the search area to a small region of interest around every cluster center, thus reducing the possibility of grouping together similar data points that belong to 2 different classes, such as the case when 2 nuclei are located close together. The general framework for segmentation and detection starts as it does in the 2D case with Gaussian de-noising and adaptive histogram equalization that enhances contrast in the volumes.

To perform supervoxel clustering, each voxel v in a volume is represented by a

vector of features $f_v = [L_v, A_v, B_v, x_v, y_v, z_v, H_v]^T$ where L_v , A_v , and B_v represent CIELAB colorspace values at v . x_v , y_v , and z_v , represent the coordinates of v within a volume.

The last component of the feature vector, H_v represents the entropy of information value at v . The entropy of a voxel is computed by calculating the entropy of the data set represented by voxel v and all its immediate neighbors using 8-connectivity and computed by the equation,

$$H(v) = \sum_{\forall n} p(i_n) \log(p(i_n)) \quad (3.1)$$

where n are all the 8-connected neighbors [52, 53] of v , i_n is the gray scale value of a neighbor, and $p(i_n)$ is the probability of that value in the neighborhood. Entropy was added to the feature vector representing a voxel for the added reliability it provides for distinguishing edges during clustering. Entropy can represent variability in information [66], and, as such, it has high energy in image regions near natural edges. It thus adds more similarity to data points coming from the same sources, background or cell nuclei, and by extension more reliability in that the clusters or supervoxels produced, will perform well in separating image objects.

Like in the 2D case, clustering was initialized by distributing cluster centers at uniform spatial intervals in the volume; the values of the feature vectors f_v at each cluster center served as the initial centers of those supervoxels. The algorithm searches for potential data points to associate with each cluster center based on the distance

between the data point and the cluster center. Distance was measured using equations 3.2 and 3.3, where m is defined exactly as it was defined in the 2D case, subsection 2.2.1.

$$d_{lab} = \sqrt{\frac{(L_{v1} - L_{v2})^2 + (A_{v1} - A_{v2})^2 + (B_{v1} - B_{v2})^2}{m}} \quad (3.2a)$$

$$d_{xyz} = \sqrt{(x_{v1} - x_{v2})^2 + (y_{v1} - y_{v2})^2 + (z_{v1} - z_{v2})^2} \quad (3.2b)$$

$$d_H = H_{v1} - H_{v2} \quad (3.2c)$$

$$d = d_{lab} + d_{xyz} + d_h \quad (3.3)$$

The search region of interest is limited to the cube whose vertices are the immediate neighboring cluster centers and the cluster center of interest at its center. For each iteration, the algorithm computes distances from one data point to all potential neighboring cluster centers, e.g., 26 potential clusters in the case of a 3D volume, and assigns it to the closest cluster. At the end of an iteration, a new cluster center for each supervoxel is computed as the mean of all voxels assigned to it. Clustering terminates when the residual error computed as the sum of all L1 distances between new cluster centers and old cluster centers is below a preset threshold, details of the algorithm were already described in subsection 2.2.1.

Figure 3.4 shows the clustering result, with one particular nucleus in focus, where we aim to demonstrate the ability of the clustering algorithm to produce clusters (supervoxels), that enclose nuclei, both capturing the shape in individual slices, and enclosing objects that traverse multiple slices. In the first slice the nucleus in question is not visible and as we move through the volume, we first slice through the upper boundary of the supervoxel enclosing the nucleus, represented by the big red blob in slice 3. After which, as the slices move into the nucleus, we noticed the shape of the supervoxel boundary adhere to the nucleus boundaries, it starts shrinking again as the slices start exiting the nucleus. Finally, both the nucleus and the supervoxel enclosing it disappear after slice 16.

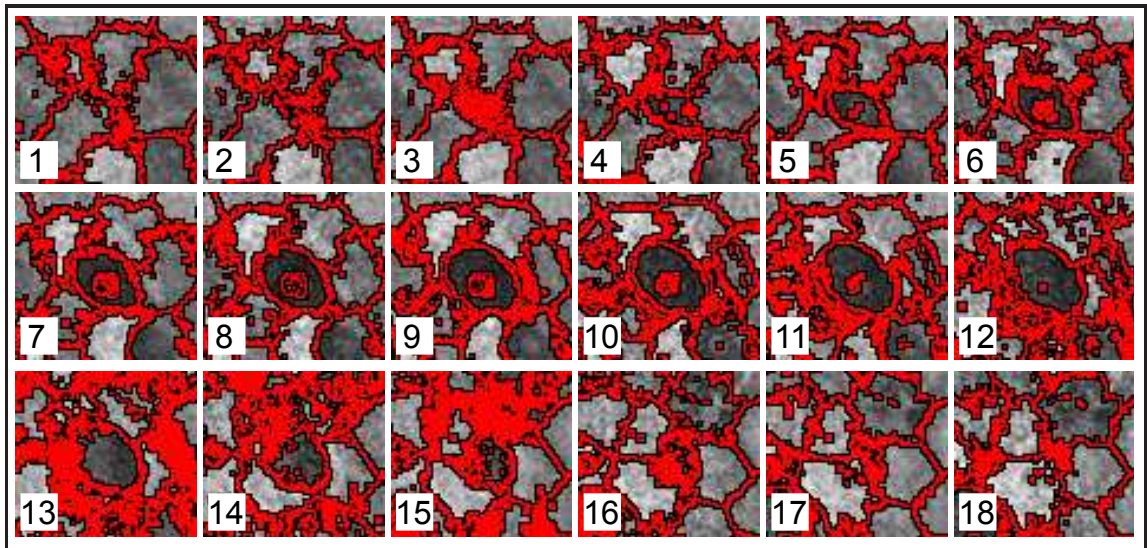


Figure 3.4 : Consecutive slices through a CARS volume showing a single nucleus being enclosed in a single supervoxel across all slices

Following this process, we move to solving the second part of the problem, computing features that can describe the physical characteristics of the supervoxels, in

order to determine those of them that enclose nuclei, as opposed to those enclosing background. This is the subject of section 3.3.

3.3 Nuclei Detection From Supervoxels

Following the generation of supervoxels, the segmentation can be completed by determining which supervoxels correspond to cells as opposed to background, thus once again the problem becomes a pattern recognition problem. We aim to design features that can provide a physical description of the clusters, in addition to representing the context which the cells are perceived through. The features were computed and used to train a semi-supervised learning machine, that classified clusters as nuclei or non nuclei.

3.3.1 Cluster Describing Features

To characterize the morphological features of supervoxels, a set of 11 features was designed for individual supervoxels aiming to represent their intensity and shape information. Specifically, and for the intensity, the *mean*, *median*, *minimum*, *maximum*, and *standard deviation* in each supervoxel was computed. To describe the clusters in context, we compared each supervoxels' intensity median and mean to those of its immediate neighboring supervoxels. Then an additional 4 features were extracted as minimums and means of the above comparisons as described by algorithm 3. Comparing features captures the subtle intensity variations between nuclei and neighboring

background supervoxels.

Algorithm 3 Computing supervoxel context indices

for each supervoxel V **do**
 Locate all neighbor supervoxels V_{nj}
 compute: $\mu \left\{ \frac{\mu_{IV}}{\mu_{IV_{nj}}} \right\} \forall_j$
 compute: $\min \left\{ \frac{\mu_{IV}}{\mu_{IV_{nj}}} \right\} \forall_j$
 compute: $\mu \left\{ \frac{\text{median}_{IV}}{\mu_{IV_{nj}}} \right\} \forall_j$
 compute: $\min \left\{ \frac{\text{median}_{IV}}{\mu_{IV_{nj}}} \right\} \forall_j$
end for

Another set of features was designed to represent supervoxels shape consistency. It was observed that the shape of background supervoxels tends to be have many irregularities, mainly caused by energy from noise disrupting the clustering process, whereas cell nuclei supervoxels are roughly characterized by ellipsoidal shapes. To quantify this difference, the distribution of the lengths of radii emitted from the center of a supervoxel to the point where they intersect its surface (boundaries of the supervoxel) [54] was computed, and shape uniformity was described as the standard deviation of these lengths. We present further details of this descriptor herein.

In their most general form, rays are just straight lines at the origin in a 3D space. For our desired use, it is required that we have multiple rays starting at the center of a supervoxel, and separated by intervals, that are user controlled. It is easiest to achieve that end if the rays are represented using spherical coordinates, with the parametric equations of the form 3.4.

$$X = R. \cos \theta. \sin \phi \quad (3.4a)$$

$$Y = R. \sin \theta. \sin \phi \quad (3.4b)$$

$$Z = R. \cos \phi \quad (3.4c)$$

Where R , θ , and ϕ , are the radial distance, the inclination angle, and the azimuth angle of a spherical coordinate system respectively. The desired ray burst effect can be produced by generating such lines at the desired separation through varying θ and ϕ . We generated the rays with a step size of $\pi/4$ for each of θ and ϕ . Figure 3.5 depicts an example of rays generated in a 3D space, note that the separation used in generating this demonstrative figure was smaller than the $\pi/4$ value used for the actual feature computation.

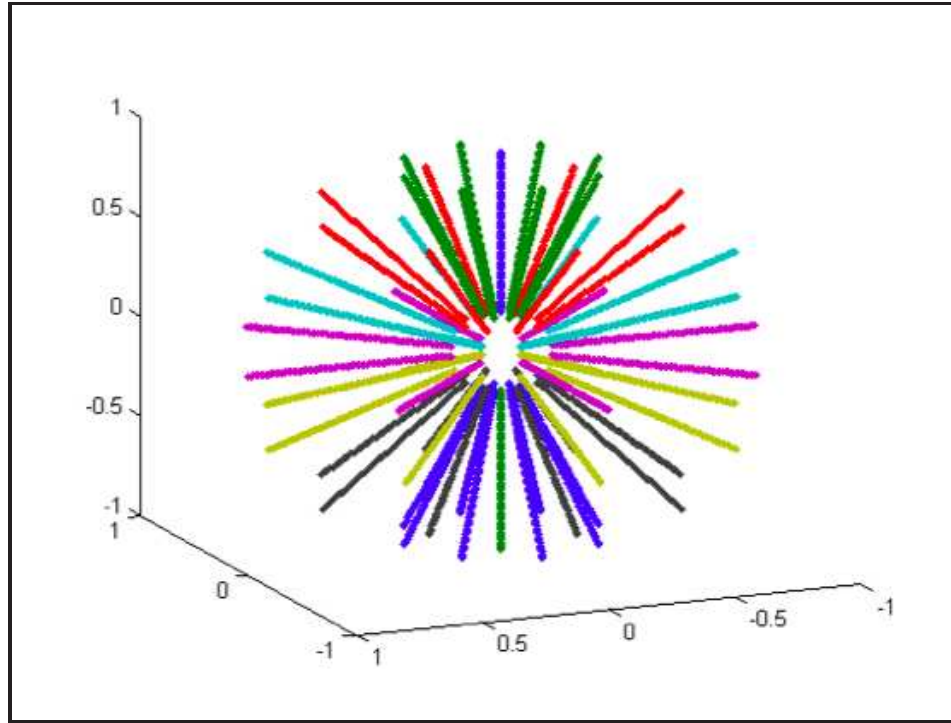


Figure 3.5 : A depiction of a general burst of rays in 3D space

The next part of computing the feature of interest is using the rays to obtain a representation of the the boundary of a supervoxel. To achieve that, first the coordinates for rays starting at the origin are computed in a general 3D space. Then, the rays are translated by the coordinates of the centroid of a supervoxel - this operation is performed for every supervoxel separately, figure 3.6-(A) depicts a rendering of a single supervoxel from a CARS volume with rays placed inside it.

What remains is finding the points of intersection between the rays and the boundary of the cluster, this is trivial once the rays are superimposed on the volume as depicted in figure 3.6-(A). Simply multiplying the 2 data sets, will reveal the intersection points depicted in figure 3.6-(B). With this information the Euclidean distance

from the cluster's centroid is computed to every point, and the feature representing the shape of the cluster is computed as the standard deviation of all these distances.

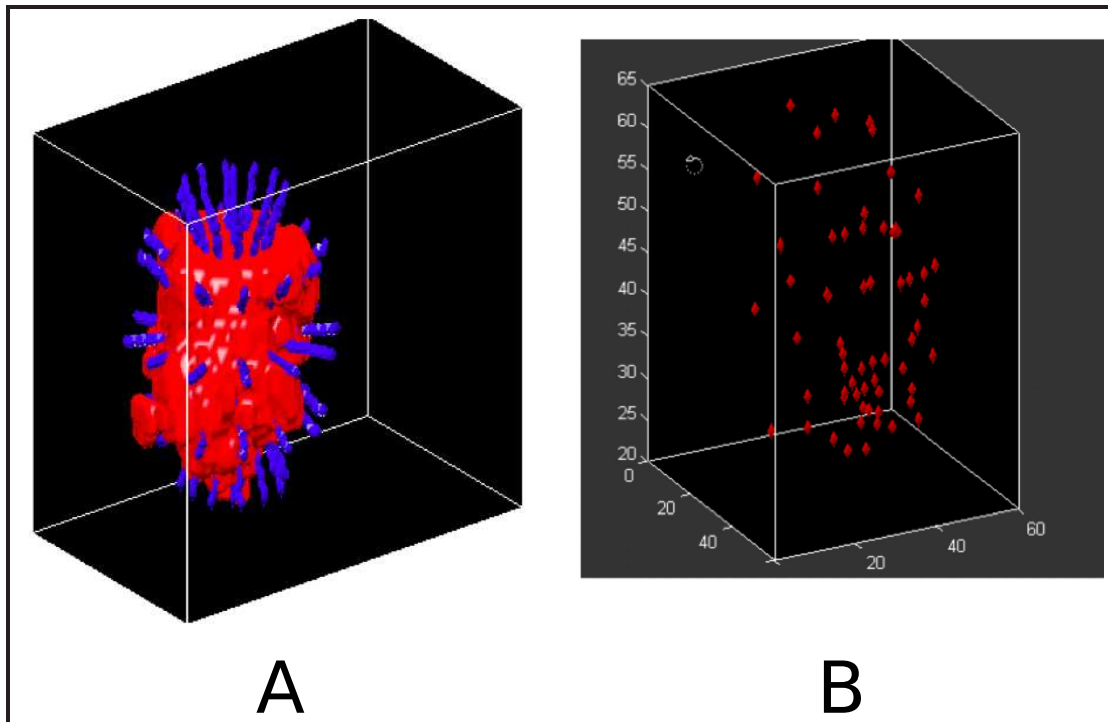


Figure 3.6 : Ray bursts inside a supervoxel (A). Intersection of rays with the supervoxel boundaries (B)

A final shape descriptor is computed as the surface area, to volume ratio of a supervoxel, it was computed for a supervoxel V using 3.5, where N is the number of voxels. The feature vector describing a voxel is of the from 3.6, Where iV is the intensity in a supervoxel - or cluster - and the rest of the terms are as defined above.

$$\frac{N_{boundaryV}}{N_V} \quad (3.5)$$

$$\left[\begin{array}{c}
 \mu_{iV} \\
 median_{iV} \\
 \min_{iV} \\
 \max_{iV} \\
 \sigma_{iV} \\
 \mu \left\{ \frac{\mu_{iV}}{\mu_{iVn_j}} \right\} \forall_j \\
 \min \left\{ \frac{\mu_{iV}}{\mu_{iVn_j}} \right\} \forall_j \\
 \mu \left\{ \frac{median_{iV}}{\mu_{iVn_j}} \right\} \forall_j \\
 \min \left\{ \frac{median_{iV}}{\mu_{iVn_j}} \right\} \forall_j \\
 \sigma_{\|rays\|} \\
 \frac{N_{boundaryV}}{N_V}
 \end{array} \right] \quad (3.6)$$

3.3.2 Nuclei Detection with Semi-Supervised Learning Machines

Following computing the quantitative features, we employ graph transduction, a semi-supervised learning approach, as the classifier to determine which supervoxels correspond to cell nuclei. Semi supervised learning is advantageous in solving problems that involve the classification of large data sets, where large amounts of labeled training data are not available, such as out CARS images. It utilizes both a small number of labeled training data points and the underlying data structure of the unlabeled data [63, 67] and is thus superior to other classifiers when the training data is limited [63, 67]. In brief, the idea of graph transduction is to discover the intrinsic man-

ifold structure underlying the data set in question by propagating on a constructed weighted graph and then to make predictions for unlabeled data. This method pursues a balance between accuracy and smoothness, that is, the classifier should make correct decisions on the previously labeled data while assigning the same labels to data points that neighbor them on the graph [67].

Mathematically graph transduction is formulated as an optimization problem. Given a data set $X = \{x_1, x_2, \dots, x_k, x_{k+1}, \dots, x_m\}$ each point of which belongs to one of the classes $C = \{1, 2, \dots, c\}$, in this case supervoxels that either belong to nuclei or to background. Few training data points are labeled, say the first k data points, x_1, x_2, \dots, x_k with labels $y_i \in C$ and the rest unlabeled. The graph transduction method predicts the labels of the unknown data points by searching for a non-negative $m \times c$ matrix F that is used to find labels for the unlabeled data points, new labels are defined as $y_i = \arg \max_{j \leq c} \{F_{ij}\}$. The matrix F is computed by optimizing the objective function 3.7 [63, 67].

$$\Phi(F) = \frac{1}{2} \left(\sum_{i,j=1}^m W_{ij} \left\| F_i / \sqrt{D_{ii}} - F_j / \sqrt{D_{jj}} \right\|^2 + \mu \sum_{i=1}^m \left\| F_i - Y_i \right\|^2 \right) \quad (3.7)$$

Where,

$$W_{ij} = \begin{cases} \exp \left(-\|x_i - x_j\|^2 / 2\sigma^2 \right) & i \neq j \\ 0 & i = j \end{cases} \quad (3.8)$$

is an $m \times m$ affinity matrix for the the data points in question, and

$$D_{ii} = \sum_{j=1}^m W_{ij} \quad (3.9)$$

is an $m \times m$ Diagonal matrix. The remaining term of the objective function in 3.10 is an $m \times c$ matrix that contains the labels y_i encoded into it's rows. For example, the first row represents the label y_1 for the first data point x_1 , if $x_1 \in \text{class 1}$, the first row Y_1 of Y will be $[1 \ 0]$, if $x_1 \in \text{class 2}$, $Y_1 = [0 \ 1]$, and so on. Rows corresponding to unlabeled data points are all zero.

$$Y_{ij} = \begin{cases} 1 & y_i = j \\ 0 & y_i \neq j \end{cases} \quad (3.10)$$

F_i and F_j are the i -th and j -th rows of F , respectively, Y_i is the i -th row of Y , and for the unlabeled data points. The optimized non-negative matrix is then used to generate the unknown labels through the rule in equation 3.11

$$y_i = \arg \max_{j \leq c} \{F_{ij}\} \quad (3.11)$$

In equation 3.7, the first term is a smoothness function that forces neighboring data points to belong to the same class. The second term is a fitting function which limits the labeled data points to conform to their original labels. The optimal F^* satisfies $F^* = \arg \max_F \{\Phi(F)\}$. To solve for F^* , we differentiate 3.7 with respect to F , and solve for F in $\frac{\partial \Phi}{\partial F} = 0$

$$\left. \frac{\partial \Phi}{\partial F} \right|_{F=F^*} = F^* - SF^* + \mu(F^* - Y) \quad (3.12)$$

then,

$$F^* - \frac{1}{1+\mu}SF^* - \frac{\mu}{1+\mu}Y = 0 \quad (3.13)$$

and,

$$F^* = \beta(1 - \alpha S)^{-1}Y \quad (3.14)$$

where,

$$S = D^{1/2}WD^{-1/2} \quad (3.15)$$

and,

$$\alpha = \frac{1}{1+\mu} \quad (3.16a)$$

$$\beta = \frac{\mu}{1+\mu} \quad (3.16b)$$

This concludes our presentation of the segmentation and detection system for 3D CARS volumes. We present and discuss experimental testing results in what follows.

3.4 Results and Discussion

3.4.1 3D Segmentation and Detection Results

We randomly selected ten image stacks from adenocarcinoma and squamous cell carcinoma, respectively, to evaluate the proposed method. It is worth noting here that automated 3D segmentation was not tested with small cell carcinoma images. As described in the introduction to this report small cell carcinoma is easily distinguishable from NSCLC using 2D data, so the need does not arise to put small cell carcinoma data through a lengthier and more resource intensive 3D segmentation and data analysis pipeline.

The average precision and recall of the proposed method compared to manual analysis were 97.8% and 92%, which indicate the robustness of the proposed method. Figure 3.7 (A-C) shows a cross section of sample CARS volume through all stages of the segmentation and detection process, A shows the original volume, the clustering result and supervoxel boundaries are presented in B, and C shows the detection result with only the boundaries of the detected supervoxels remaining. In addition, a rendering of the segmented cell nuclei is in panel D of the same figure.

Figure 3.8 shows the same nucleus we showed back in figure 3.4 after classification of supervoxels into nuclei or background. The figure shows that classification eliminates irrelevant supervoxels, and preserves those that enclose nuclei, noting that in some slices of figure 3.8 an additional nucleus is visible, causing the boundary of another supervoxel, to be visible next to the supervoxel of interest.

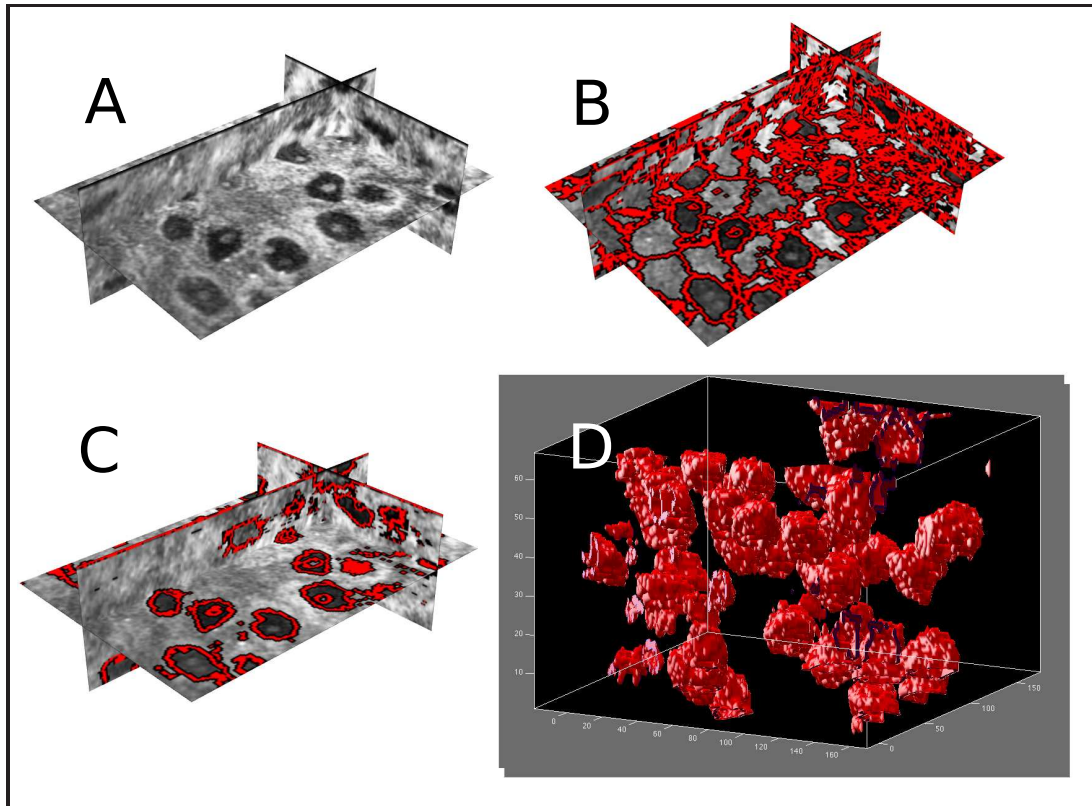


Figure 3.7 : A segmentation and detection demonstration in a 3D CARS volume, with a rendering of the segmented and detected cell nuclei

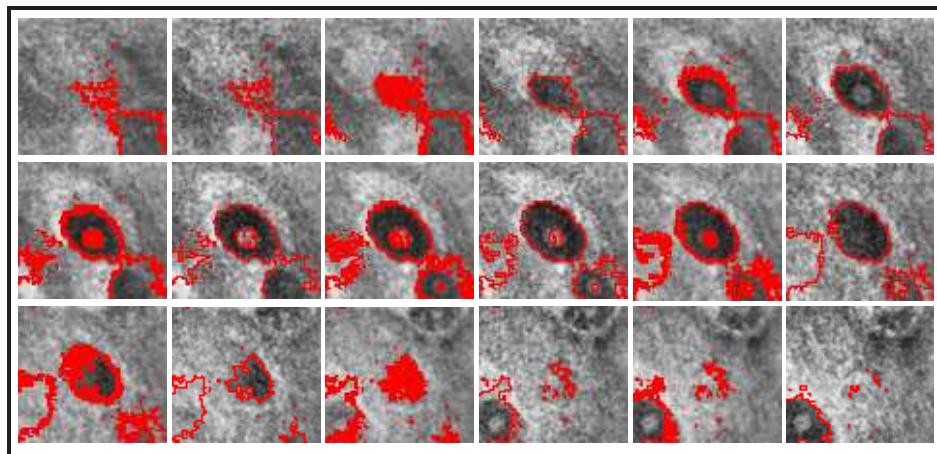


Figure 3.8 : Consecutive slices through a CARS volume - after classification - showing a nucleus being enclosed in a single supervoxel across all slices

We benchmarked our method against a set of popular cell nuclei detection and segmentation methods, performing a qualitative comparison with the iterative radial voting (IRV) for nuclei detection [68], local binary fitting (LBF) level set nuclei segmentation method designed for uneven background conditions [69], and Otsu’s thresholding method, tests clearly showed the superior nuclei detection capability of our method. In addition we performed a quantitative comparison in detection results between the graph transduction method we used for detecting nuclei and support vector machine (SVM) classification [70].

Briefly, IRV was designed to deal with noisy data and scale variation [68]. The LBF method is a region-based active contour method which aims at overcoming the intensity variations and uneven background by utilizing a local binary energy fitting function [69]. SVM is one of the most popular supervised learning methods, it searches for a hyper-plane that best separates training data through maximizing a separation margin [70].

Comparison results are provided in Figure 3.9. The IRV method detected many false positive and negative cell nuclei centers, represented by the red crosses in panel A. The LBF method incorrectly classified background regions into cell nuclei due to the low intensity and noisy signal of background regions (B). The performance of Otsu’s thresholding method (C) was worse compared against the performance of LBF. Due to the limited training data, SVM (D) mislabeled many cell nuclei and background regions, whereas our proposed method appeared to be superior to the

above mentioned methods (E). From this qualitative comparison and comparing the validation scores of graph transduction to those of SVM when classifying supervoxels to nuclei or background in table 3.1, we conclude that the proposed method, which integrates local clustering and semi-supervised learning machines, handles 3D CARS image segmentation well.

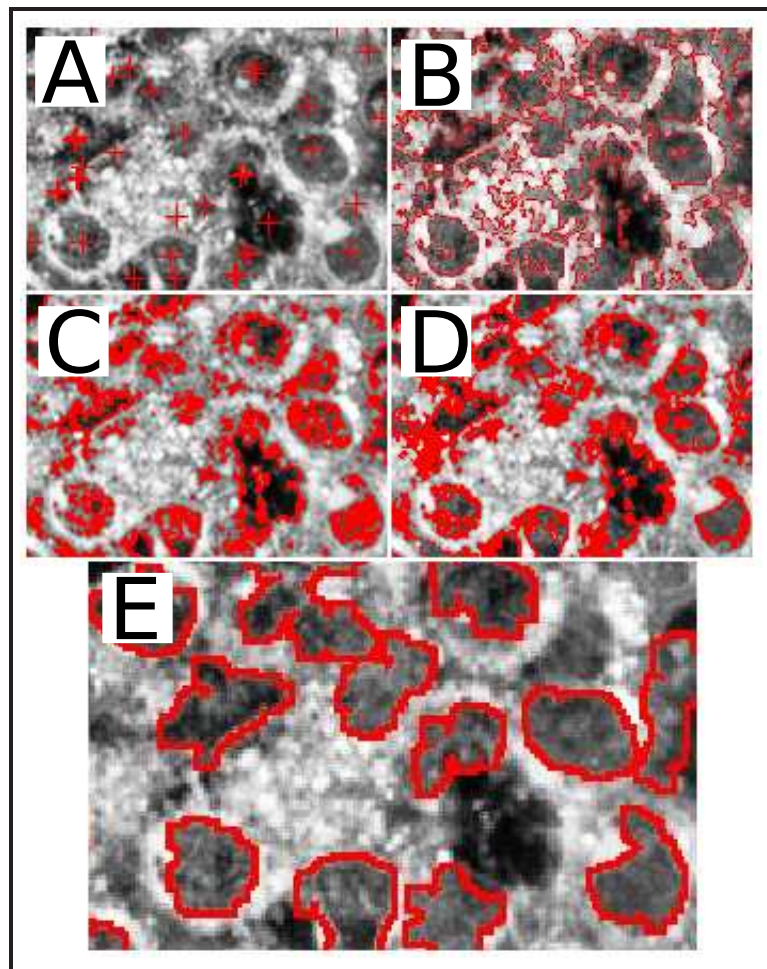


Figure 3.9 : Comparing supervoxel nuclei detection results from our method to other nuclei detection methodologies

Table 3.1 : A quantitative comparison between graph transduction and SVM quality of supervoxel nuclei detection

Method	Precision	Recall	f-Score
Proposed	98.8%	92%	94.3%
SVM	96%	79.8%	86.2%

3.4.2 Discussion

In this Chapter, we proposed a new framework to solve the 3D cellular nuclei segmentation problem in CARS image stacks, or volumes. In order to overcome the low signal-to-noise ratio and uneven background challenges in CARS images, we extended the local clustering algorithm to operate on 3D data sets. The local clustering method provides continuous partitioning regions, or supervoxels, of individual nuclei across all dimensions in a CARS volume. These regions were again are quantified and classified with graph transduction, a semi-supervised learning method. Graph transduction significantly reduces the obstacles facing a good classification, these obstacles are mainly caused by limited training data and significant intraclass variations in the training data.

Experimental results have shown that the proposed framework performed well in segmenting and detecting cellular nuclei in 3D CARS image stacks. Furthermore, the method was demonstrated to be superior to other widely used nuclei detection and segmentation approaches. It is the first automated 3D nuclei segmentation approach in CARS microscopy images, to the best of our knowledge. The developed approach enables the automatic segmentation of cell nuclei and thus paves the way for the

ultimate goal we wish attain in this study, automatic differential diagnosis of NSCLC. In the following chapter we build on the current results, by extracting pathologically relevant morphological features from the segmented nuclei and exploring how these features can be utilized in a fully automated differential diagnosis of NSCLC cell types.

Chapter 4

Automatic Differential Diagnosis of NSCLC Cell Types

4.1 introduction

Thus far, our focus has been on sifting through a very large data set of three-dimensional images to find a small subset of data that contains information relevant to diagnosis, namely those voxels corresponding to cell nuclei. The end goal however remains as from the beginning, the development of a system for automatic differential diagnosis of non-small cell lung carcinoma cell types. As stated in chapter 1, the motivation behind developing a fully automated differential diagnosis platform for NSCLC is acquiring the ability to rapidly recognize different cancer cell types, with minimal tissue consumption, which in turn would accelerate cancer diagnosis,

This chapter is adapted from the following publications:

1. Liang Gao, Ahmad Hammoudi, Fuhai Li, Michael J. Thrall, Philip Cagle, Yuanxin Chen, Jian Yang, XiaofenXia, Yubo Fan, Yehia Massoud, Zhiyong Wang, Stephen T.C. Wong, "Differential Diagnosis of Lung Carcinoma with Three-dimensional Quantitative Molecular Vibrational Imaging". Currently in submission to the International Society for Optics and Photonics Journal of Biomedical Optics.
2. Liang Gao, Ahmad Hammoudi, Fuhai Li, Michael J. Thrall, Philip Cagle, Yuanxin Chen, Jian Yang, XiaofenXia, Yubo Fan, Yehia Massoud, Zhiyong Wang, Stephen T.C. Wong, "Label-free Diagnosis of Lung Cancer Subtypes with Three-dimensional Molecular Vibrational Imaging". Abstract 7890, the American Association for Cancer Research Annual Meeting, March 31 - April 4, 2012. Chicago, IL.

preserve tissue samples for subsequent molecular testing, and aid in delivering better targeted therapy.

What we have developed and detailed in the previous chapters, is a fully automated system capable of detecting, segmenting, and computing all relevant physical information about the geometry and spatial distribution of cellular nuclei in label-free CARS images. In this chapter, we explain how the information extracted by the detection and segmentation system, in other words the subset of relevant information within the large CARS data set, we explain how it was utilized to compute physical features of the nuclei, and the tissue as a whole, and how these information were in turn used to perform automatic differential diagnosis.

The features we aim to design and compute in this chapter are different from all the features that were discussed in the chapters 2, and 3 in both the characteristics we try to capture with them, and their purpose, it is worth noting the differences to avoid any confusion. Specifically, in the earlier chapters, the features represented superpixels or supervoxels, which were just clusters of data points in space. In this chapter however, the features were used to represent cellular nuclei. Second, the information underlying the features and their utilization are also different in this chapter. In the previous case, our purpose was to determine whether the clusters represented by the features were nuclei, or not. Now, only nuclei are studied, and the features we to compute are used to determine what type of NSCLC nuclei are they, adenocarcinoma, or squamous cell carcinoma, thus a different set of features is

required.

So, to achieve the end goal of automatic differential diagnosis of NSCLC cell types, we designed a new set of features that we believe capture morphological characteristics of the nuclei, and spatial characteristics of the cancerous tissue as a whole that carry pathologically relevant information. Information that would allow us to build a classification system that delivers differential diagnostic accuracy of NSCLC cell unattainable with current state of the art automatic diagnosis methods, and thus could enhance the quality of diagnosis and targeted therapy delivery to lung carcinoma patients . First, in section 4.2 we try to unify the physical description of the nuclei by fitting 3D ellipsoids to the points in space representing each nucleus. Then, we build set of feature describing those ellipsoids, and also the spatial patterns they create in the imaged tissue, this is the subject of section 4.3. Finally, and in section 4.4 we use these features to build a classifier that can determine to what cell type do the cells in a tissue of interest belong to. We conclude the chapter with a brief discussion in section 4.5.

4.2 Ellipsoid Fitting

To ensure consistency in extracted features across different volumes, and because different tissue samples were imaged at different times and having different freshness, we assumed a cell nucleus to be roughly ellipsoidal in shape, and, as such, an ellipsoid was fitted to each contiguous set of 3D points representing a nucleus. The features

we designed for each nucleus later on were the physical features of the ellipsoid that best fitted the convex hull of the points comprising that nucleus, in other words, the ellipsoid that best fitted the smallest region in space that could enclose all the points comprising a nucleus. This is called the minimum volume enclosing ellipsoid or MVEE its full details were presented in [71, 72]. We describe it briefly in what follows. Figure 4.1 shows a rendering of a segmented CARS volume, in which we use the points from each rendered nucleus to fit an ellipsoid.

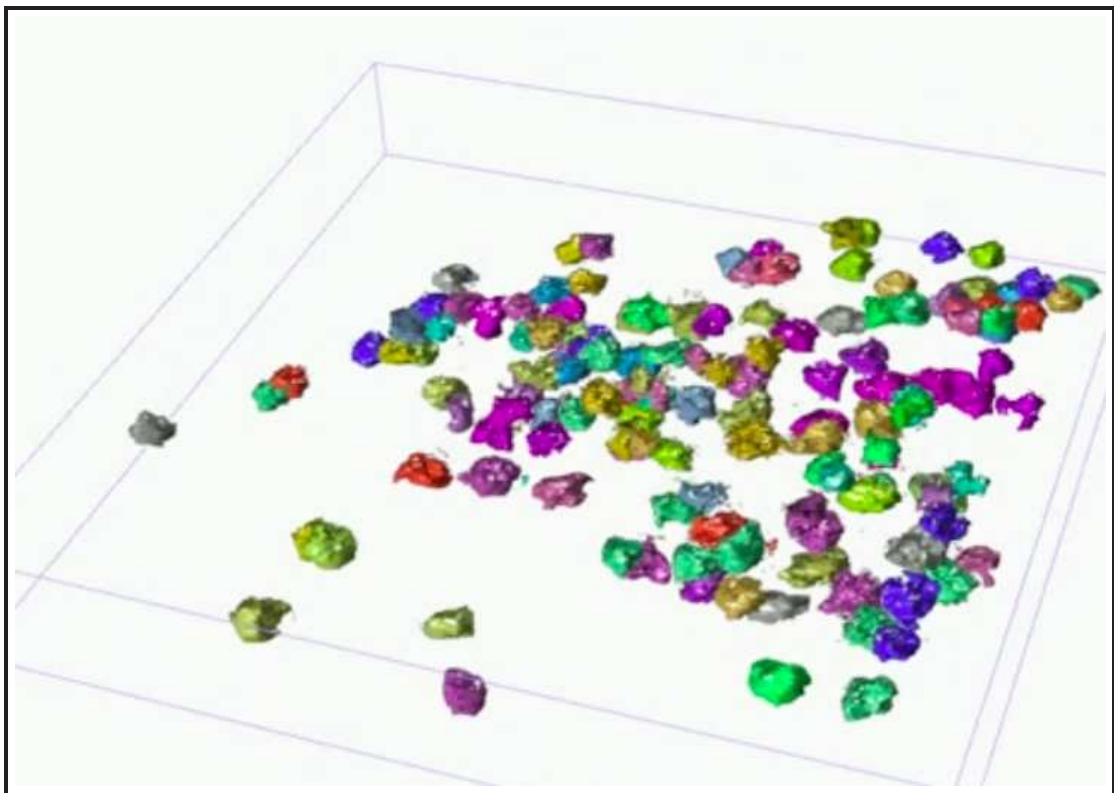


Figure 4.1 : A rendering of a full volume of segmented cell nuclei

To perform Ellipsoid fitting, we solve the optimization problem described in [72] and represented in 4.1

$$\begin{aligned}
& \underset{A,c}{\text{minimize}} && \log\{\det(A)\} \\
& \text{subject to} && (P_i - c)^T A^* (P_i - c) \leq 1
\end{aligned} \tag{4.1}$$

Where A is a symmetric positive definite matrix that represents all the information about an ellipsoid, c is a vector representing the coordinates of the center of the ellipsoid, and P_i is a vector containing the i^{th} column of the matrix P . Finally P is a $3 \times N$ matrix, having the coordinates of all the points being fit - in the case the points in a nucleus - as columns, and N is the number of such points. Many solvers are available for the minimization problem we used one based on the Kachiyan algorithm [73] to determine A and c , from which all information about a given ellipsoid can be computed as follows.

The Matrix A and center c , which are computed as the solution to the optimization program in 4.1 are used to represent a 3D ellipsoid in centric form shown in 4.2.

$$\varepsilon = \{x \in \mathbb{R}^n \mid (x - c)^T A^{-1} (x - c) = 1\} \tag{4.2}$$

This representation of an Ellipsoid can be used to render the Ellipsoid in space, and compute it's relevant parameters such as it's volume, surface area, and lengths of it's equatorial radii, we will come to discuss those in section 4.3 as they will be used to compute the pathologically relevant features. For now, we show how a CARS volume looks after fitting ellipsoids to each of the segmented nuclei, a single nucleus with the corresponding MVEE is presented in figure 4.2, the blue points represent the

nucleus, and the mesh the MVEE. While a full volume with each nucleus fitted to its MVEE is presented in figure 4.3, the color variation in that figure represents depth within the tissue, with blue representing the highest part of the tissue, and red the deepest, the mesh is representative of the real size and distribution of nuclei within the imaged tissue.

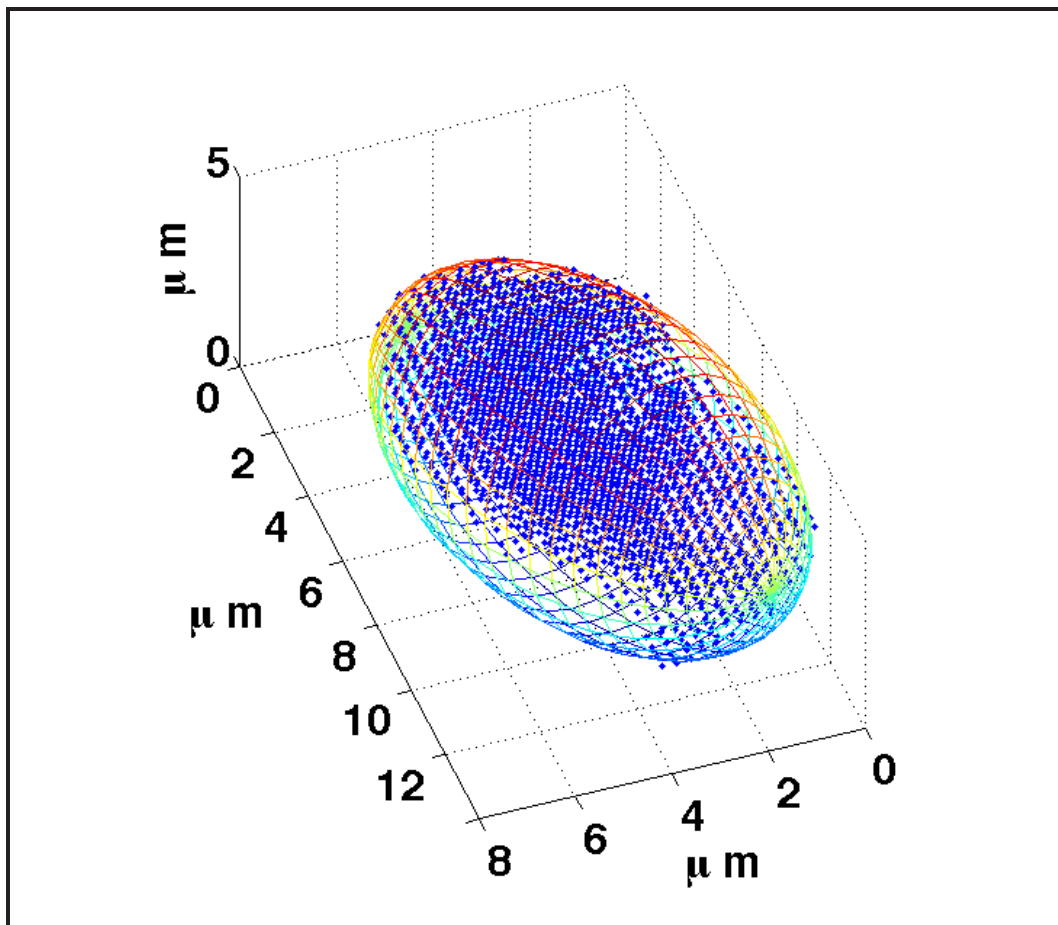


Figure 4.2 : A single nucleus (Blue), and the corresponding MVEE (Mesh)

In what follows, we present how the MVEEs we fitted to our CARS volume were used to extract pathologically relevant features that were used in building an auto-

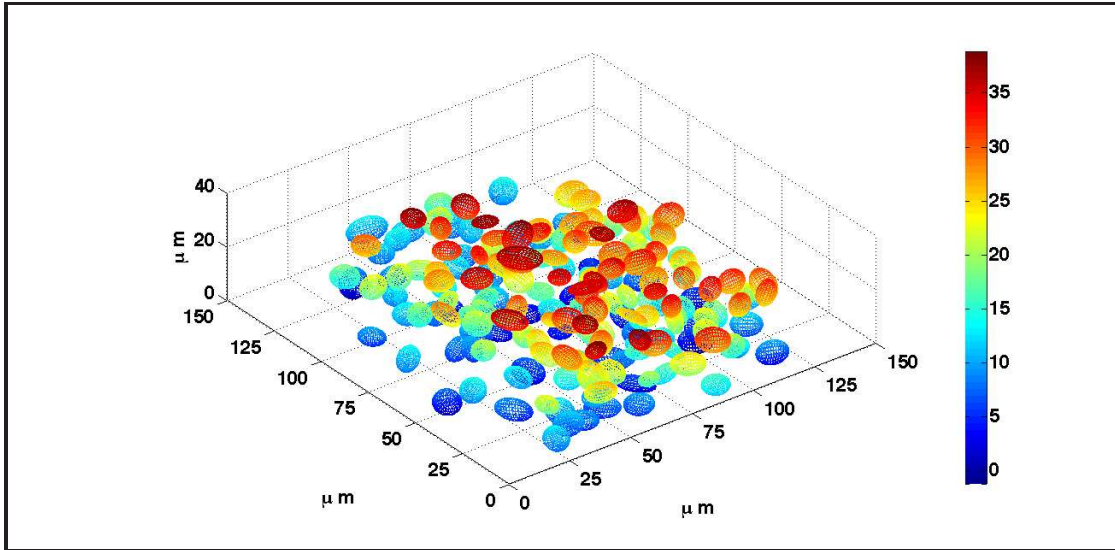


Figure 4.3 : MVEEs for all nuclei in a full CARS volume

mated differential classifier.

4.3 Designing Pathologically Relevant Features

A series of pathologically-related morphological features such as nuclear size, and cell-to-cell distances have been previously used for disease identification, and providing meaningful diagnostic information with reproducible results. These features have been previously tested with various disease models such as lung, breast and prostate cancers [13, 14, 25–27]. Herein, we compute a set of such morphological features from 3D CARS volumes that we believe will provide meaningful information for the differential diagnosis of NSCLC cell types.

4.3.1 Single Nucleus Features

We compute two sets of features, one related directly to each nucleus on its own, and another set of features that aims to describe the distribution of the nuclei in the tissue, and their relation to one another. we begin with the cell related features.

As we mentioned in section 4.2, all the information about an ellipsoid can be extracted from the matrix A and the center c of equation 4.1. The square roots of the eigenvalues σ_A of A represent the lengths of the equatorial radii of the ellipsoid, or its major axis, and its two minor axes. The eigenvectors of A represent the direction vectors of the ellipsoid, those are shown in figure 4.4.

With this information, and alternate representation of an ellipsoid is available, and has the form in 4.3

$$\varepsilon = \left\{ \frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1 \right\} \quad (4.3)$$

Where x , y , and z are the coordinates of the center of the Ellipsoid, and a , b , and c are the lengths of the equatorial radii. We use this definition of an ellipsoid to compute the following features for each nucleus.

- *Volume* of the Ellipsoid given by $\frac{4}{3}\pi abc$
- *Length* of the major axis, or first equatorial radius, given by a of equation 4.3
- *Lengths* of the second and third equatorial radii, or the two minor axes of the ellipsoid, given by b and c of equation 4.3

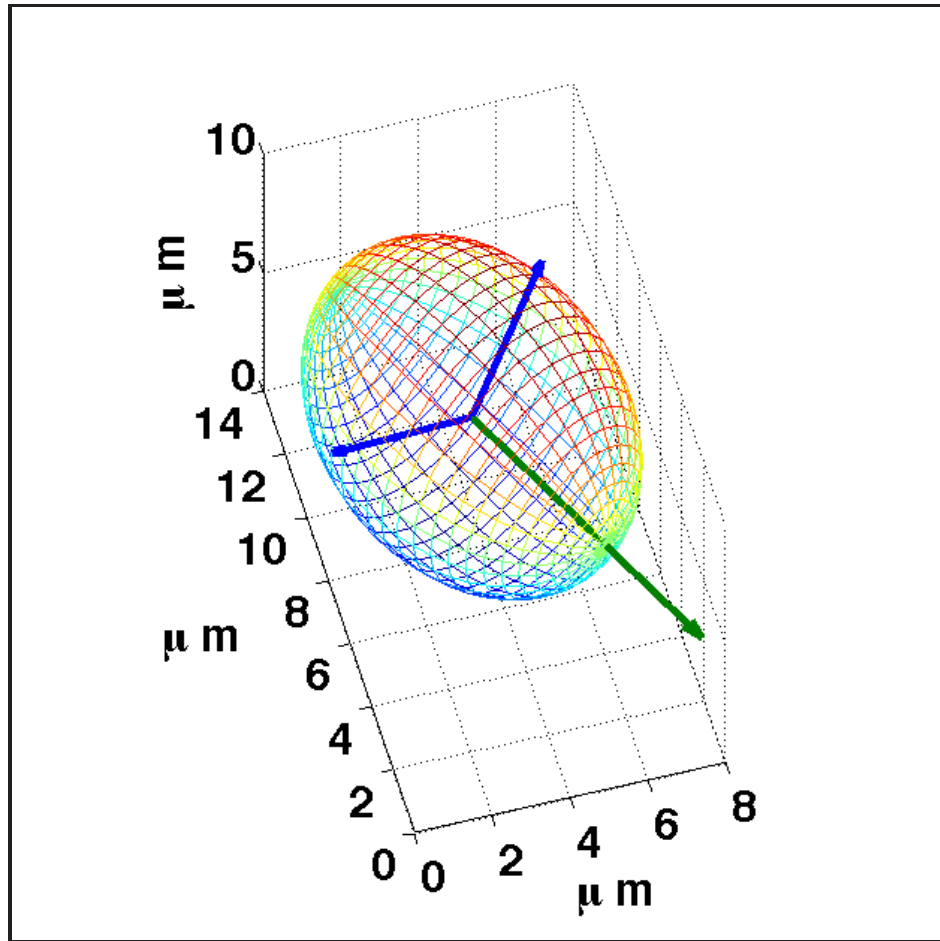


Figure 4.4 : An ellipsoid with its 3 direction vectors shown

4.3.2 Whole Tissue Features

Furthermore, to compute features related to a nucleus relative to its neighbors, we constructed the Delaunay triangulation [74–76] to define the connectivity of nuclei in the sample.

The Delaunay triangulation of a set of points is the division of the region of space containing these points into simplexes or tetrahedra with the points of interest at their vertices, this defines a graph as that presented in figure 4.5, the edges of the

tetrahedra - which are also the edges of the graph - connect all the nuclei, in this case the graph vertices. The neighbors of a nucleus of interest were assumed to be those nuclei joined to it by an edge of the triangulation. Using the triangulation, four features were computed for each nucleus to describe its spatial configuration with respect to the rest of the nuclei and the tissue.

The features were distance to farthest neighbor, distance to nearest neighbor, mean distance to neighbors, and orientation relative to nearest neighbor, the features are listed in detail below.

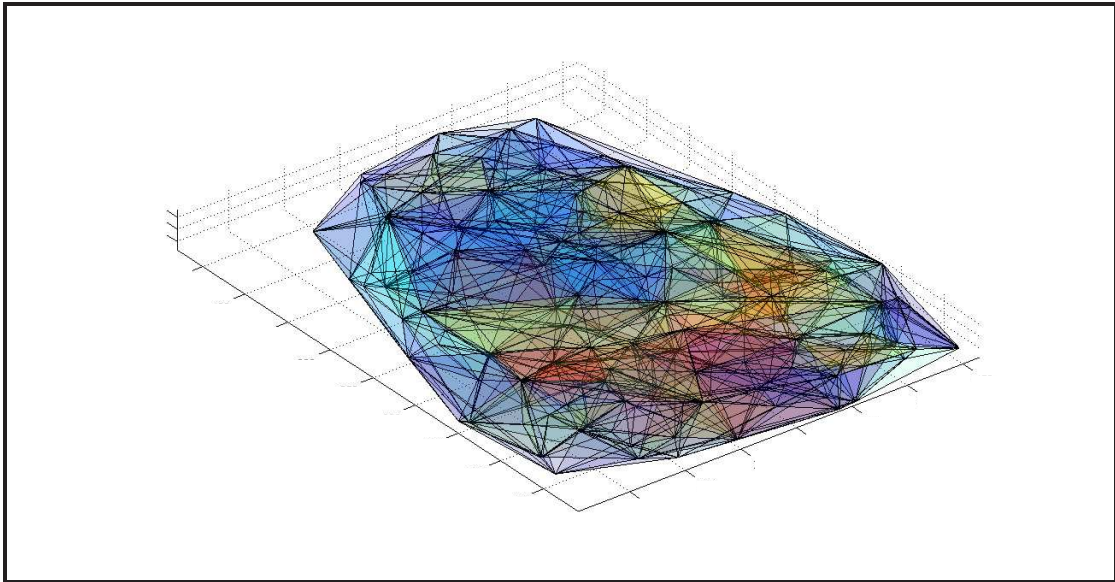


Figure 4.5 : Delaunay triangulation of nuclei in a CARS image

- Distance to farthest neighbor is computed as the length of the longest edge of the Delaunay triangulation graph attached to the nucleus of interest, given by

$$\max_{\forall i} \{\|i\|_2\} \quad (4.4)$$

where i is a Delaunay edge connected to the nucleus.

- Distance to nearest neighbor: The length of the shortest Delaunay edge attached to the nucleus, or

$$\min_{\forall i} \{\|i\|_2\} \quad (4.5)$$

- Mean distance to neighbors, N is the number of neighbors, also given by the number of Delaunay edges connected to a nucleus

$$\frac{\sum_{\forall i} \{\|i\|_2\}}{N_i} \quad (4.6)$$

- Orientation relative to the nearest neighbor is assumed to be the angle between the major axis of a cell and the major axis of its nearest neighbor, is computed from the dot product between the direction vector along the major axis of the cell and that of its nearest neighbor.

$$\cos^{-1} \langle a_v, a_{vn} \rangle \quad (4.7)$$

Where a_v is the direction vector along the major axis of the nucleus of interest,

and a_{vn} is the direction vector along the the major axis of its nearest neighbor.

In summary, for each of our CARS volume, we computed features relating to the size each nucleus, namely the volume, and length of its equatorial radii. In addition, we computed features that present some information about how nuclei in a volume are related, those were derived from a Delaunay triangulation of all nuclei in a volume, and include cell-cell distance, and orientation of nuclei relative to one another.

In section 4.4 we take a closer look at the computed features and their statistics. In addition, we design a method of using those statistics in determining whether the cells comprising a tissue in a CARS volume belongs to the adenocarcinoma or squamous cell carcinoma cell types .

4.4 Differential Classification

We performed segmentation, labeling, and feature extraction from 15 volumes corresponding to adenocarcinoma and another 15 volumes from squamous cell carcinoma. In addition, because a main goal of our work is demonstrating that analyzing CARS data in a higher dimension than previously explored will solve the NSLC differential diagnosis problem, we benchmarked against 2D based differential diagnosis. We extracted three slices from each of the 15 image stacks and computed the same nuclear features from those as 2D images. It is worth noting that all features can be mapped to any dimension, in which an ellipsoid mapped to two dimensions becomes an ellipse or volume becomes area. The only major difference between 3D and 2D representa-

tion of the features is that an ellipse has only two equatorial radii composed of one major and one minor axis, as opposed to an ellipsoid which has three such radii. As such, 3D measurements resulted in one more feature over 2D, as an additional minor axis length since there are three axes in a 3D ellipsoid.

Since each volume contains hundreds of nuclei, and since diagnosis aims to classify an entire tissue to a specific NSCLC cell type, as opposed to classifying individual cells, we use the statistics of the computed features to perform the classification. In other words, for each of the volumes, all the nuclei are segmented, and the features described in section 4.3 are computed for each nucleus, then the probability distribution functions (PDF) of each of the features across all nuclei in a image stack are used to describe that image stack. So a volume is identified by the statistics of the nuclear features rather than individual features of nuclei. Figures 4.6 and 4.7 show the PDF's of each of the features for our entire data set, in addition to a comparison between 3D features, and 2D features.

We observed that 5 out of the 8 measured features showed significant separation or difference between 2D and 3D measurements. They were mostly related to the size, shape and orientation of nuclear structures, figure 4.6. We note here that features related to nuclear size and shape failed to show a clear separation between the two cancer subtypes in 2D measurements, figure 4.6 (A-C). In contrast, 3D measurements effectively captured the difference between the two cell types by showing clearly separate peaks of the PDF curves, figure 4.6 (E-G). While nuclear orientation shows a

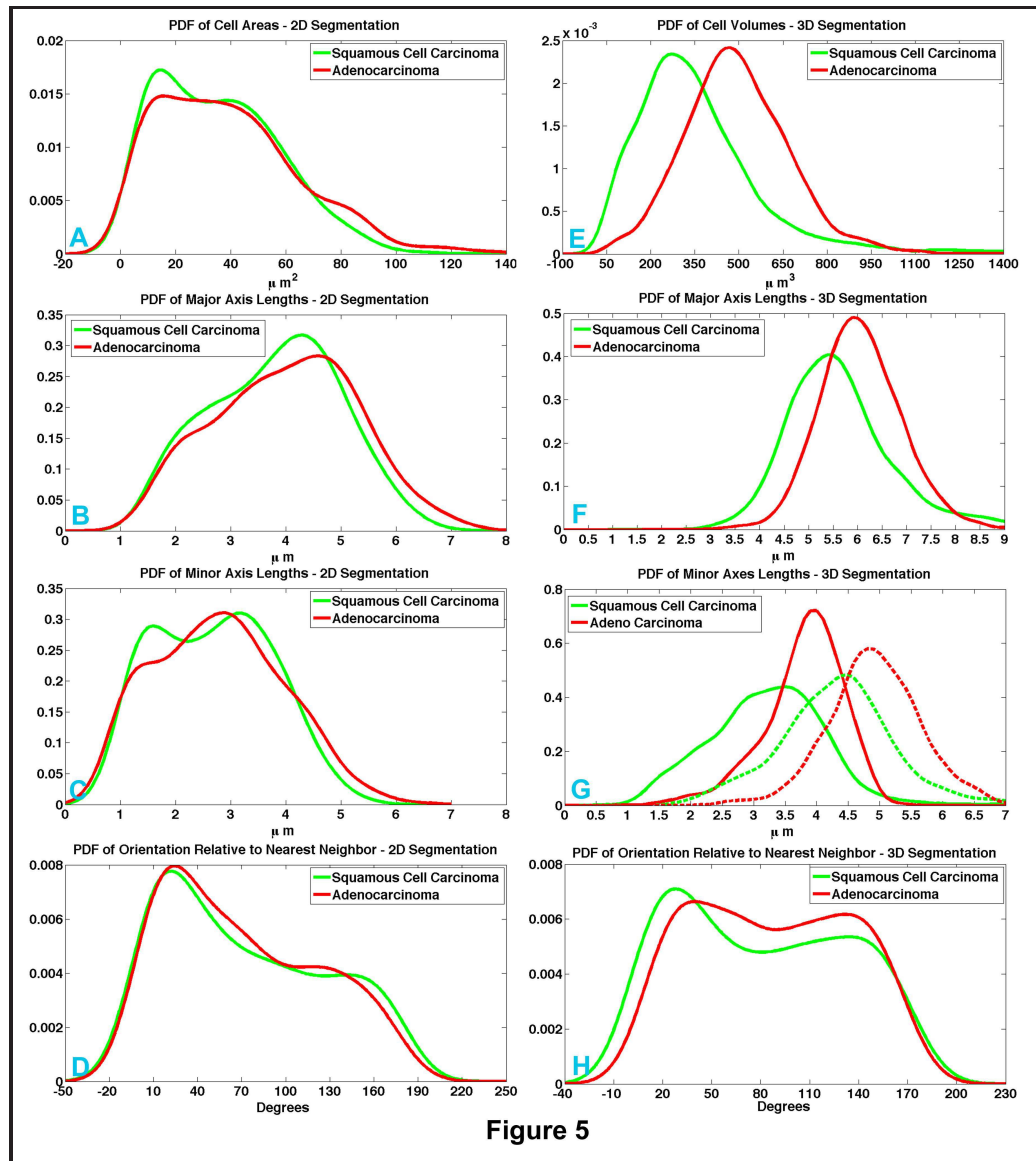


Figure 4.6 : Statistics of single nucleus derived features - 2D vs. 3D

clear peak with 2D measurement, figure 4.6 (D), this peak turns into a more uniform distribution across different angles in 3D, figure 4.6 (5). The presented PDF curves were estimated from the measured data, i.e., they represent the histograms of the features processed with a smoothing kernel causing the curves in some cases to have

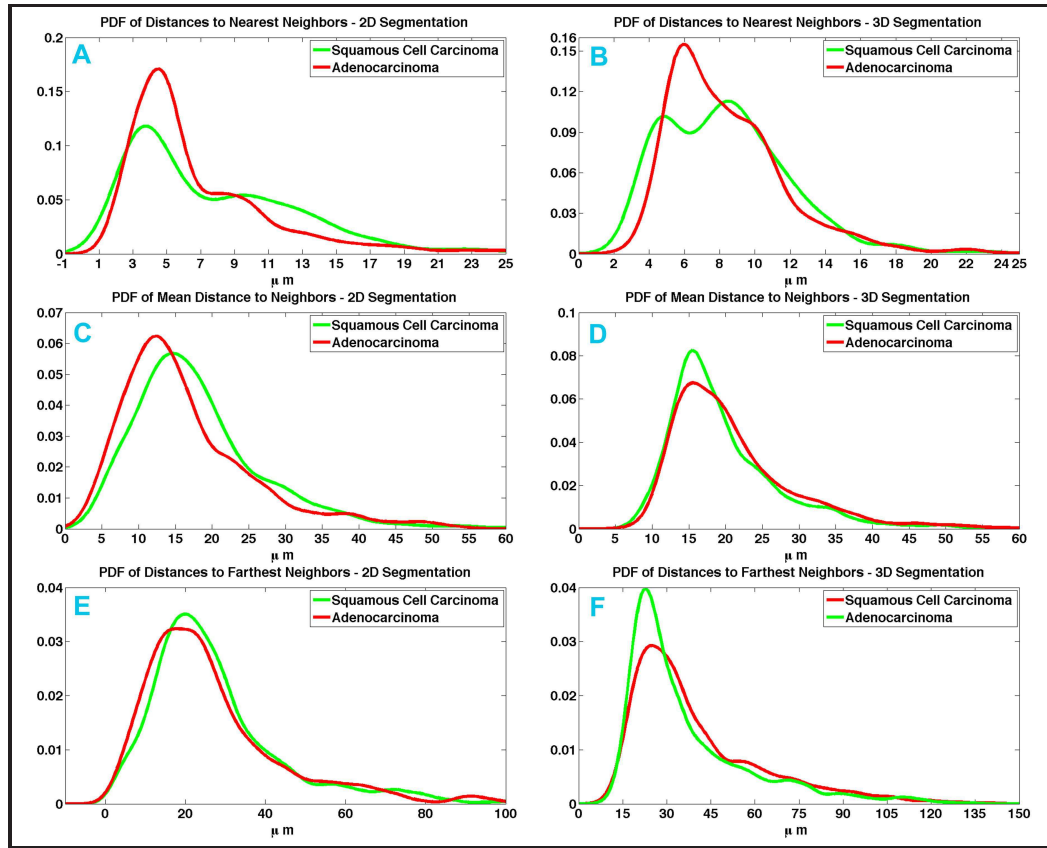


Figure 4.7 : Statistics of tissue derived features - 2D vs. 3D

tails in the negative side of the number line.

Moreover, and since a volume is identified by the statistics of its features we describe those statistics by computing the mean, standard deviation, skewness, and kurtosis of the distribution function for each feature. i.e. for the seven 2D features and for the 8 3D features in a given volume.

These quantities constitute the 32 components of the feature space used to build a classifier able to characterize a volume as either an adenocarcinoma or squamous cell carcinoma cell types.

We employed, the leave-one-out cross validation approach, also used for differential diagnosis of lung carcinoma subtypes by [13], to demonstrate that the designated statistical features could be used to perform automatic diagnosis. We use the same classification method used in prior work [13] to validate our hypothesis that studying features in 3D space, and that our set of features is in fact superior to prior methods when used in diagnosis. Moreover we do this to demonstrate that the method of examining features and the features themselves are the real bottleneck to achieving high quality differential diagnosis, regardless of the classification method used.

The leave-one-out classification works as follow; over a sufficient number of iterations, one sample of each adenocarcinoma and squamous cell carcinoma was excluded, and the remaining 28 samples, 14 of each subtype, were used to train an artificial neural network. The two left out samples were then used to test the performance of the neural network, with the same architecture described in section 2.3.2 of chapter 2 and 20 neurons in the hidden layer. This was repeated until each sample of each subtype was left out and paired with all samples from the other subtype. For example, If samples from each of the two cancer subtypes were labeled 1 through 15, A1 A15 for adenocarcinoma samples and S1-S15 for squamous cell carcinoma samples, samples A1 and S1 would be excluded on the first iteration. The remaining 28 samples would be used to train a classifier, in this case a neural network. The 2 excluded samples (A1 and S1) are used to validate that classifier through testing whether it can assign them to their correct cell type. On the second iteration, sample A1 would be left out

from adenocarcinoma with sample S2 instead of S1 from squamous cell carcinoma, and the training/validation process is repeated. This process is repeated over enough iterations to exhaust all possible pairings of samples. As a benchmark in this study, the process was performed both in 2D and in 3D data sets.

We present the differential diagnosis results in section 4.5

4.5 Results

Using the calculated features, a classifier was built, as described in section 4.4, classification was performed to separate cancer cell types. Figures 4.8 and 4.9 illustrate the automatic classification results from 2D data and 3D data respectively. In Figure 4.8, results of classification using 2-D data are presented, where each point represents one tissue sample - or volume - specifically the one that was left out during a particular iteration. The threshold for classification is the straight line $y = 0.5$. Points with $y \geq 0.5$ were classified as adenocarcinoma, while points with $y < 0.5$ were classified as squamous cell carcinoma. For better presentation, the graph is separated into two separate subfigures, respectively representing adenocarcinoma samples and squamous cell carcinoma samples. Classification from 2D data resulted in a true positive rate of 71.98% and 65.05% for adenocarcinoma and squamous cell carcinoma, respectively. False positive rate, in the same order, was 28.02% and 34.95%, table 4.1. Classification results from the 3D data analysis are plotted in figure 4.9, where clear separation between subtypes allows for data visualization on the same graph. As de-

tailed in table 4.2, quantified classification accuracies were a 99.56% and 97.78% true positive rate for adenocarcinoma and squamous cell carcinoma, respectively. False positive classification was 0.45% and 2.22% in the same order. The results showed that features extracted from 3D data analysis provided information that significantly enhanced the classification accuracy and thus demonstrated proof of concept that 3D image analysis allows for the automatic diagnosis of lung cancer cell types, with a quality far superior to differential diagnosis from 2D data.

We further discuss these results in section 4.6

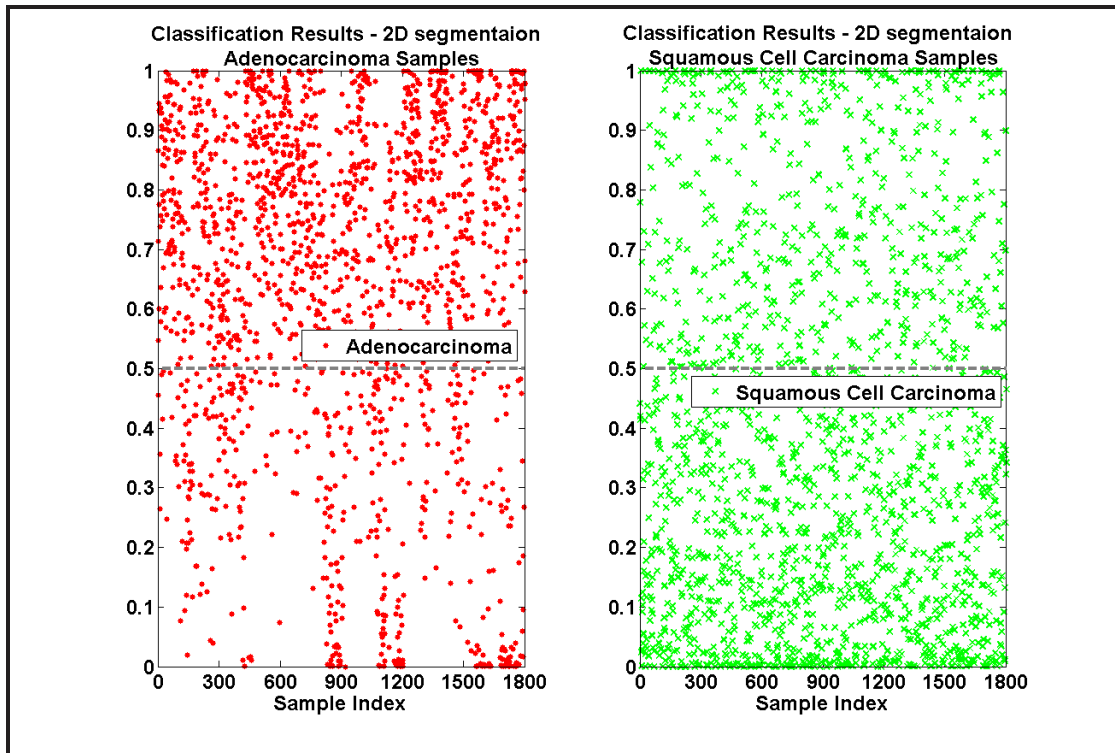


Figure 4.8 : Differential diagnosis results from 2D data

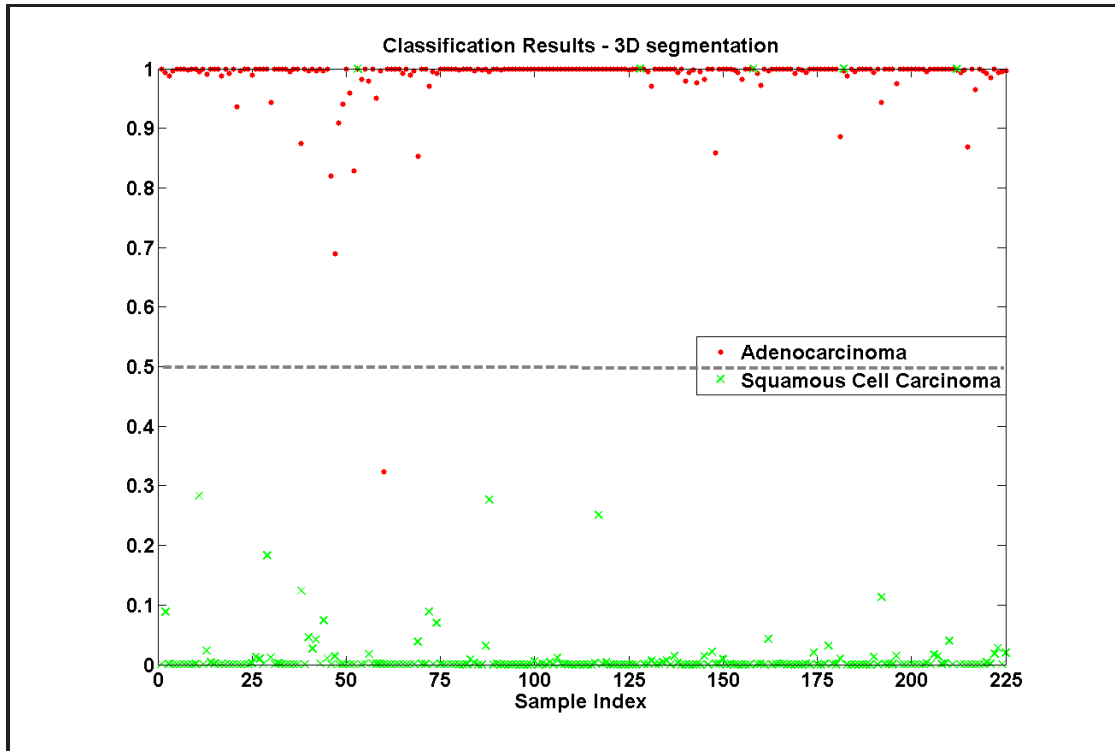


Figure 4.9 : Differential diagnosis results from 3D data

Table 4.1 : Differential diagnosis accuracies from 2D data

Truth \ Classification	Adenocarcinoma	Squamous cell carcinoma
Adenocarcinoma	71.98%	34.94%
Squamous cell carcinoma	28.02%	65.05%

Table 4.2 : Differential diagnosis accuracies from 3D data

Truth \ Classification	Adenocarcinoma	Squamous cell carcinoma
Adenocarcinoma	99.56%	0.45%
Squamous cell carcinoma	2.22%	97.78%

4.6 Discussion

In this chapter we have proposed an effective strategy to increase the accuracy for characterization of NSCLC cell types by quantifying pathological characteristics of nuclei using 3D data analysis. By taking advantage of the optical sectioning capability readily provided by CARS, a valuable advantage that is not available with H&E staining and imaging technique, our strategy allows analysis of cellular features in 3D, a setting close to the microenvironment in which the cells reside. As illustrated in Figure 4.6, measurements on 5 out of the 8 disease-related features showed clear features separation in 3D measurements, as opposed compared to 2D controls.

First, the volumes of nuclei were measured in 3D to provide size information. This effectively allowed for the suppression of the sampling error posed by measurements from 2D slices. As a result, a clear separation of adenocarcinoma from squamous cell carcinoma was observed, this was noticeable as two very separate peaks of the PDF curves. In contrast, the 2D control showed much broader distributions for both subtypes without clear separation of peak positions, this is potentially caused by multiple measurements corresponding to a single nucleus across multiple slices, thus hiding the subtle difference in peak positions.

Second, the sampling noise in 2D also obscured effective identification of the difference in nuclear size between examined cell types. It is worth mentioning that the peaks of the major axis lengths of adenocarcinoma and squamous cell carcinoma were about $6 \mu m$ and $5.5 \mu m$, respectively. Although they were well separated in 3D mea-

surements, 2D measurements of the same parameter did not show a clear difference. More importantly, both cell types showed the peak value of major axis length to fall between 4 and 5 μm , values that were substantially smaller than those measured in 3D.

These data indicates that measurements in 2D combined values associated with both sampling error, from 3D to 3D, and real axis lengths. Specifically, if, in a given plane, we assume that all cell nuclei have the same axis length, then the measured axis length for a given nucleus can range from zero to its actual length, depending on where a given slice passes through a nucleus of interest. This effect is true for every cell nucleus in that plane, resulting in a distribution of the measured value from zero to the actual length, especially when the spatial locations of individual nuclei are unrelated. As such, a sampling error is created and could result in a more uniform distribution curve that does not reveal the real feature statistics. In contrast, 3-D measurement of the same parameter resulted in much narrower PDF curves with clearly separate peaks. The same observation was supported by minor axis lengths, as well.

Third, the relative orientation angle between neighbors showed a peak around 25 degrees in 2D. This peak was not present in the 3D measurements, where a more uniform distribution was observed across the spectrum of all angles, indicating a lack of dominant orientation of nuclei orientation in such tumors. This distribution could be caused by the use of animal models, instead of human tumor samples,

as it is well known that adenocarcinomas tend to form glandular structures, while squamous cell carcinomas tend to form well-oriented cell sheets when they are well differentiated (43, 44). Consequently, by showing more random cellular distribution patterns, the mouse tumor model could be more poorly differentiated compared to human tumors. Moreover, this difference also supports the utility of the 3D approach since this method even worked well on mouse models, which are harder to separate. However, the tremendous heterogeneity of human tumors, as compared with only two cell lines in this study, may pose challenges for classification that are difficult to capture in this model system. Finally, compared to the aforementioned features, a clear distinction between 2D and 3D approaches could not be found relative to measurements of distance-related features. This was expected because the sampling noise error by 2D measurements will not affect the position of nuclei centers, which are the major determinants of cell-cell distance.

Finally, we have demonstrated that properly selected image features, and through measuring those features in an appropriate environment, namely a 3D setting that most accurately mimics the real spatial configuration of cancerous tissue. It is possible to perform fully automatic differential diagnosis of NSCLC cell types. As opposed to performing this process from 2D images that do not sample sufficient information from the cancerous tissue to allow differential classification. Our work paves a way towards a crucial advancement in lung cancer discovery and diagnosis that could greatly enhance care delivery through more accurate targeted therapy, and thus en-

hance patient survival rates.

Chapter 5

Conclusions and Future Work

Recent advances in targeted therapy hold the promise for the delivery of better, more effective treatments to lung cancer patients, that could significantly enhance their survival rates. Optimizing care delivery through target therapies requires the ability to effectively identify and diagnose lung cancer along with the specific cell type to each patient.

Label free optical imaging techniques such as CARS have the potential to provide physicians with minimally invasive access to lung tumor sites and thus allow for better cancer diagnosis and sub-typing. To maximize the benefits of such imaging modalities in enhancing cancer treatment, the development of new data analysis methods that can rapidly and accurately analyze the new types of data provided through them is essential. Recent studies have gone long ways to achieving those goals but faced some significant bottlenecks hindering the ability to fully exploit the diagnostic potential of CARS images, namely, the streamlining of the diagnosis process was hindered by the lack of ability to automatically detect cancer cell, and the inability to completely classify them into their respective cell types.

In this study we have addressed the two bottlenecks named above, through designing an image processing framework that is capable of automatically and with great

accuracy detecting cancer cells in two and three dimensional CARS images. Moreover, we built upon this capability with a new approach at analyzing the segmented data, that provided significant information about the cancerous tissue and ultimately allowed for the automatic differential classification of non-small cell lung carcinoma cell types.

From this here on, what is required to deliver a better diagnostic platform is two fold. First, and on the imaging side, transforming the CARS microscope from a table top setup to a portable endoscope is essential for the ability to use it on a mass scale with cancer patients. Second, and on the data analysis side, The image processing and classification algorithms need to be trained with a larger number of human cancer data to further assure of their dependability, they need to be optimized to be able to handle real time data, and deliver on-the-spot diagnosis, and they need to be transferred into portable hardware that can be used with CARS endoscopes.

Bibliography

- [1] D. M. Parkin, F. Bray, J. Ferlay, and P. Pisani, “Global cancer statistics, 2002,” *CA: A Cancer Journal for Clinicians*, vol. 55, no. 2, pp. 74–108, 2005.
- [2] H. Hashizume, P. Baluk, S. Morikawa, J. W. McLean, G. Thurston, S. Roberge, R. K. Jain, and D. M. McDonald, “Openings between defective endothelial cells explain tumor vessel leakiness,” *The American Journal of Pathology*, vol. 156, no. 4, pp. 1363 – 1380, 2000.
- [3] D. R. Youlton, S. M. Cramb, and P. D. Baade, “The international epidemiology of lung cancer: Geographical distribution and secular trends,” *Journal of Thoracic Oncology*, vol. 3, pp. 819 – 831, August 2008.
- [4] S. Diederich *et al.*, “Lung cancer screening: status in 2007,” *Der Radiologe*, vol. 48, no. 1, p. 39, 2008.
- [5] C. I. Henschke, D. F. Yankelevitz, D. M. Libby, M. W. Pasmantier, J. P. Smith, and O. S. Miettinen, “Survival of patients with stage I lung cancer detected on CT screening,” *The New England Journal of Medicine*, vol. 355, no. 17, pp. 1763–1771, 2006.
- [6] A. McWilliams, C. MacAulay, A. F. Gazdar, and S. Lam, “Innovative molec-

- ular and imaging approaches for the detection of lung cancer and its precursor lesions,” *Oncogene*, vol. 21, pp. 6949–6959, Oct. 2002.
- [7] P. Cagle, T. Allen, S. Dacic, M. Beasley, A. Borczuk, L. Chirieac, R. Laucirica, J. Ro, and K. Kerr, “Revolution in lung cancer: new challenges for the surgical pathologist,” *Archives of Pathology & Laboratory Medicine*, vol. 135, no. 1, pp. 110–116, 2011.
- [8] M. D. Duncan, J. Reintjes, and T. J. Manuccia, “Scanning coherent anti-Stokes Raman microscope,” *Opt. Lett.*, vol. 7, pp. 350–352, Aug 1982.
- [9] C. L. Evans and X. S. Xie, “Coherent anti-Stokes Raman scattering microscopy: Chemical imaging for biology and medicine,” *Annual Review of Analytical Chemistry 2008*, vol. 1, no. 1, pp. 883–909, 2008.
- [10] J.-X. Cheng and S. Xie, “Coherent anti-Stokes Raman scattering microscopy: instrumentation, theory, and applications,” *The Journal of Physical Chemistry B*, vol. 108, no. 3, pp. 827–840, 2004.
- [11] C. L. Evans, E. O. Potma, and X. S. Xie, “Coherent anti-Stokes Raman scattering spectral interferometry: determination of the real and imaginary components of nonlinear susceptibility $\chi(3)$ for vibrational microscopy,” *Opt. Lett.*, vol. 29, pp. 2923–2925, Dec 2004.
- [12] C. Evans, E. Potma, M. Puoris’ haag, D. Côté, C. Lin, and X. Xie, “Chemical imaging of tissue in vivo with video-rate coherent anti-Stokes Raman scatter-

- ing microscopy,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46, p. 16807, 2005.
- [13] L. Gao, F. Li, M. J. Thrall, Y. Yang, J. Xing, A. A. Hammoudi, H. Zhao, Y. Massoud, P. T. Cagle, Y. Fan, K. K. Wong, Z. Wang, and S. T. C. Wong, “On-the-spot lung cancer differential diagnosis by label-free, molecular vibrational imaging and knowledge-based classification,” *Journal of Biomedical Optics*, vol. 16, no. 9, p. 096004, 2011.
- [14] Y. Yang, F. Li, L. Gao, Z. Wang, M. J. Thrall, S. S. Shen, K. K. Wong, and S. T. C. Wong, “Differential diagnosis of breast cancer using quantitative, label-free and molecular vibrational imaging,” *Biomed. Opt. Express*, vol. 2, pp. 2160–2174, Aug 2011.
- [15] M. Maemondo, A. Inoue, K. Kobayashi, S. Sugawara, S. Oizumi, H. Isobe, A. Gemma, M. Harada, H. Yoshizawa, I. Kinoshita, *et al.*, “Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR,” *New England Journal of Medicine*, vol. 362, no. 25, pp. 2380–2388, 2010.
- [16] T. Mitsudomi, S. Morita, Y. Yatabe, S. Negoro, I. Okamoto, J. Tsurutani, T. Seto, M. Satouchi, H. Tada, T. Hirashima, *et al.*, “Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (wjtog3405): an open label, randomised phase 3 trial,” *The Lancet-Oncology*, vol. 11, no. 2, pp. 121–128, 2010.

- [17] T. Mok, Y. Wu, S. Thongprasert, C. Yang, D. Chu, N. Saijo, P. Sunpaweravong, B. Han, B. Margono, Y. Ichinose, *et al.*, “Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma,” *New England Journal of Medicine*, vol. 361, no. 10, pp. 947–957, 2009.
- [18] M. Cohen, J. Gootenberg, P. Keegan, and R. Pazdur, “FDA drug approval summary: bevacizumab (avastin®) plus carboplatin and paclitaxel as first-line treatment of advanced/metastatic recurrent nonsquamous non-small cell lung cancer,” *The Oncologist*, vol. 12, no. 6, pp. 713–718, 2007.
- [19] D. Johnson, L. Fehrenbacher, W. Novotny, R. Herbst, J. Nemunaitis, D. Jablons, C. Langer, R. DeVore III, J. Gaudreault, L. Damico, *et al.*, “Randomized phase ii trial comparing bevacizumab plus carboplatin and paclitaxel with carboplatin and paclitaxel alone in previously untreated locally advanced or metastatic non-small-cell lung cancer,” *Journal of Clinical Oncology*, vol. 22, no. 11, pp. 2184–2191, 2004.
- [20] P. Loo, S. Thomas, M. Nicolson, M. Fyfe, and K. Kerr, “Subtyping of undifferentiated non-small cell carcinomas in bronchial biopsy specimens,” *Journal of Thoracic Oncology*, vol. 5, no. 4, p. 442, 2010.
- [21] W. Travis, E. Brambilla, M. Noguchi, A. Nicholson, K. Geisinger, Y. Yatabe, C. Powell, D. Beer, G. Riely, K. Garg, *et al.*, “International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory

- Society: International multidisciplinary classification of lung adenocarcinoma: Executive summary,” in *Proceedings of the American Thoracic Society*, vol. 8, p. 381, Am Thoracic Soc, 2011.
- [22] C. Henschke, D. McCauley, D. Yankelevitz, D. Naidich, G. McGuinness, O. Miettinen, D. Libby, M. Pasmantier, J. Koizumi, N. Altorki, *et al.*, “Early lung cancer action project: overall design and findings from baseline screening,” *The Lancet*, vol. 354, no. 9173, pp. 99–105, 1999.
- [23] F. Levi, F. Lucchini, E. Negri, and C. La Vecchia, “Trends in mortality from major cancers in the european union, including acceding countries, in 2004,” *Cancer*, vol. 101, no. 12, pp. 2843–2850, 2004.
- [24] M. Duncan, J. Reintjes, and T. Manuccia, “Imaging biological compounds using the coherent anti-Stokes Raman scattering microscope,” *Optical Engineering*, vol. 24, no. 2, pp. 352–355, 1985.
- [25] L. Gao, Y. Yang, J. Xing, M. J. Thrall, Z. Wang, F. Li, P. Luo, K. K. Wong, H. Zhao, and S. T. C. Wong, “Diagnosing lung cancer using coherent anti-Stokes Raman scattering microscopy,” in *Proceedings of SPIE* (A. Mahadevan-Jansen, T. Vo-Dinh, and W. S. Grundfest, eds.), vol. 7890, p. 789015, SPIE, 2011.
- [26] L. Gao, H. Zhou, M. J. Thrall, F. Li, Y. Yang, Z. Wang, P. Luo, K. K. Wong, G. S. Palapattu, and S. T. C. Wong, “Label-free high-resolution imaging of

- prostate glands and cavernous nerves using coherent anti-Stokes Raman scattering microscopy,” *Biomed. Opt. Express*, vol. 2, pp. 915–926, Apr 2011.
- [27] Y. Yang, L. Gao, Z. Wang, M. J. Thrall, P. Luo, K. K. Wong, and S. T. Wong, “Label-free imaging of human breast tissues using coherent anti-Stokes Raman scattering microscopy,” in *Proceedings of SPIE* (A. Periasamy, K. König, and P. T. C. So, eds.), vol. 7903, p. 79032G, SPIE, 2011.
- [28] L. Gao, Y. Yang, J. Xing, M. Thrall, Z. Wang, F. Li, P. Luo, K. Wong, and S. Wong, “Differential diagnosis of human lung cancer x2014; a label-free and chemistry-sensitive approach,” in *Life Science Systems and Applications Workshop (LiSSA), 2011 IEEE/NIH*, pp. 14–17, april 2011.
- [29] J. Terry, S. Leung, J. Laskin, K. Leslie, A. Gown, and D. Ionescu, “Optimal immunohistochemical markers for distinguishing lung adenocarcinomas from squamous cell carcinomas in small tumor samples,” *The American Journal of Surgical Pathology*, vol. 34, no. 12, p. 1805, 2010.
- [30] C. Langer, B. Besse, A. Gualberto, E. Brambilla, and J. Soria, “The evolving role of histology in the management of advanced non–small-cell lung cancer,” *Journal of Clinical Oncology*, vol. 28, no. 36, p. 5311, 2010.
- [31] A. Hammoudi, F. Li, L. Gao, Z. Wang, M. Thrall, Y. Massoud, and S. Wong, “Automated nuclear segmentation of coherent anti-Stokes Raman scattering microscopy images by coupling superpixel context information with artificial neural

- networks,” *Machine Learning in Medical Imaging*, vol. 7009, pp. 317–325, 2011.
- [32] N. Otsu, “A threshold selection method from gray-level histograms,” *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.
- [33] R. F. Moghaddam and M. Cheriet, “Adotsu: An adaptive and parameterless generalization of otsu’s method for document image binarization,” *Pattern Recognition*, vol. 45, no. 6, pp. 2419–2431, 2011.
- [34] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, “Global cancer statistics,” *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.
- [35] L. Shafarenko, M. Petrou, and J. Kittler, “Automatic watershed segmentation of randomly textured color images,” *IEEE Transactions on Image Processing*, vol. 6, pp. 1530–1544, Nov 1997.
- [36] X. Yang, H. Li, and X. Zhou, “Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and kalman filter in time-lapse microscopy,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, pp. 2405–2414, Nov. 2006.
- [37] V. Caselles, R. Kimmel, and G. Sapiro, “Geodesic active contours,” *International Journal of Computer Vision*, vol. 22, no. 1, pp. 61–79, 1997.
- [38] Y. Boykov, O. Veksler, and R. Zabih, “Fast approximate energy minimization via

- graph cuts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1222–1239, Nov 2001.
- [39] C. CHEN, H. LI, X. ZHOU, and S. T. C. WONG, “Constraint factor graph cut-based active contour method for automated cellular image segmentation in RNAi screening,” *Journal of Microscopy*, vol. 230, no. 2, pp. 177–191, 2008.
- [40] T. Chan and L. Vese, “Active contours without edges,” *IEEE Transactions on Image Processing*, vol. 10, pp. 266–277, Feb 2001.
- [41] R. Goldenberg, R. Kimmel, E. Rivlin, and M. Rudzsky, “Fast geodesic active contours,” *IEEE Transactions on Image Processing*, vol. 10, pp. 1467–1475, Oct 2001.
- [42] Z. Yin, R. Bise, M. Chen, and T. Kanade, “Cell segmentation in microscopy imagery using a bag of local bayesian classifiers,” in *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 125–128, April 2010.
- [43] A. Lucchi, K. Smith, R. Achanta, V. Lepetit, and P. Fua, “A fully automated approach to segmentation of irregularly shaped cellular structures in EM images,” *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*, pp. 463–471, 2010.
- [44] X. Ren and J. Malik, “Learning a classification model for segmentation,” in

- Ninth IEEE International Conference on Computer Vision Proceedings*, pp. 10–17 vol.1, oct. 2003.
- [45] J. A. Hartigan and M. A. Wong, “Algorithm as 136: A k-means clustering algorithm,” *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. pp. 100–108, 1979.
- [46] A. Likas, N. Vlassis, and J. J. Verbeek, “The global k-means clustering algorithm,” *Pattern Recognition*, vol. 36, no. 2, pp. 451 – 461, 2003.
- [47] T. Moon, “The expectation-maximization algorithm,” *IEEE Signal Processing Magazine*, vol. 13, pp. 47 –60, Nov 1996.
- [48] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, “SLIC superpixels,” *Technical Report 149300 EPFL*, June 2010.
- [49] A. Levinstein, A. Stere, K. Kutulakos, D. Fleet, S. Dickinson, and K. Siddiqi, “Turbopixels: Fast superpixels using geometric flows,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 2290 –2297, Dec. 2009.
- [50] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, “Adaptive histogram equalization and its variations,” *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355 – 368, 1987.
- [51] K. Zuiderveld, *Contrast limited adaptive histogram equalization*, pp. 474–485.

San Diego, CA, USA: Academic Press Professional, Inc., 1994.

- [52] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, pp. 182 – 214, 601 – 615. Prentice Hall Professional Technical Reference, 2002.
- [53] R. Gonzalez and E. Richard, Woods, *Digital Image Processing*, pp. 75 – 147. Prentice Hall Press, 2002.
- [54] K. Smith, A. Carleton, and V. Lepetit, “Fast ray features for learning irregular shapes,” in *IEEE 12th International Conference on Computer Vision*, pp. 397 –404, Oct 2009.
- [55] T. Mitchell, *Machine learning*, ch. 4, pp. 81–124. McGraw Hill, 1997.
- [56] M. H. Beale, M. T. Hagan, and H. B. Demuth, *Matlab Neural Network Toolbox User Guide*, March 2012.
- [57] P. Sahoo, S. Soltani, and A. Wong, “A survey of thresholding techniques,” *Computer Vision, Graphics, and Image Processing*, vol. 41, no. 2, pp. 233 – 260, 1988.
- [58] S. Beucher, “The watershed transformation applied to image segmentation,” *Scanning Microscopy International*, vol. 6, pp. 299–314, 1992.
- [59] X. Zhou, K. Y. Liu, P. Bradley, N. Perrimon, and S. T. Wong, “Towards automated cellular image segmentation for rnai genome-wide screening,” in *Med-*

- ical Image Computing and Computer-Assisted Intervention – MICCAI 2005*, vol. 3749 of *Lecture Notes in Computer Science*, pp. 885–892, 2005.
- [60] T. R. Jones, A. Carpenter, and P. Golland, “Voronoi-based segmentation of cells on image manifolds,” in *Computer Vision for Biomedical Image Applications*, vol. 3765 of *Lecture Notes in Computer Science*, pp. 535–543, 2005.
- [61] C. Li, C. Xu, C. Gui, and M. Fox, “Level set evolution without re-initialization: a new variational formulation,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 430 – 436 vol. 1, June 2005.
- [62] H. Liu, Y. Chen, H. P. Ho, and P. Shi, “Geodesic active contours with adaptive neighboring influence,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005*, vol. 3750 of *Lecture Notes in Computer Science*, pp. 741–748, 2005.
- [63] J. Wang, T. Jebara, and S.-F. Chang, “Graph transduction via alternating minimization,” in *Proceedings of the 25th International Conference on Machine Learning, ICML '08*, (New York, NY, USA), pp. 1144–1151, ACM, 2008.
- [64] O. Veksler, Y. Boykov, and P. Mehrani, “Superpixels and supervoxels in an energy optimization framework,” in *Computer Vision – ECCV 2010*, vol. 6315 of *Lecture Notes in Computer Science*, pp. 211–224, 2010.

- [65] A. Lucchi, K. Smith, R. Achanta, G. Knott, and P. Fua, “Supervoxel-Based segmentation of mitochondria in EM image stacks with learned shape features.,” *IEEE Transactions on Medical Imaging*, 2011.
- [66] P. S. Shenkin, B. Erman, and L. D. Mastrandrea, “Information-theoretical entropy as a measure of sequence variability,” *Proteins: Structure, Function, and Bioinformatics*, vol. 11, no. 4, pp. 297–313, 1991.
- [67] X. Zhu, “Semi-supervised learning literature survey,” 2005.
- [68] B. Parvin, Q. Yang, J. Han, H. Chang, B. Rydberg, and M. Barcellos-Hoff, “Iterative voting for inference of structural saliency and characterization of sub-cellular events,” *IEEE Transactions on Image Processing*, vol. 16, pp. 615–623, March 2007.
- [69] C. Li, C.-Y. Kao, J. Gore, and Z. Ding, “Implicit active contours driven by local binary fitting energy,” in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–7, June 2007.
- [70] C. J. Burges, “A tutorial on support vector machines for pattern recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
- [71] P. Kumar and E. A. Yildirim, “Minimum-volume enclosing ellipsoids and core sets,” *Journal of Optimization Theory and Applications*, vol. 126, no. 1, pp. 1–21, 2005.

- [72] N. Moshtagh, “Minimum volume enclosing ellipsoid,” *Convex Optimization*, 2005.
- [73] P. Gács and L. Lovász, “Khachiyan’s algorithm for linear programming,” in *Mathematical Programming at Oberwolfach*, vol. 14 of *Mathematical Programming Studies*, pp. 61–68, 1981.
- [74] Y. Xu, L. Liu, C. Gotsman, and S. J. Gortler, “Capacity-constrained Delaunay triangulation for point distributions,” *Computers and Graphics*, vol. 35, no. 3, pp. 510 – 516, 2011.
- [75] F. Li, X. Zhou, J. Ma, and S. Wong, “Multiple nuclei tracking using integer programming for quantitative cancer cell cycle analysis,” *IEEE Transactions on Medical Imaging*, vol. 29, pp. 96 –105, Jan. 2010.
- [76] F. Li, X. Zhou, and S. T. C. Wong, “Optimal live cell tracking for cell cycle study using time-lapse fluorescent microscopy images,” in *Machine Learning in Medical Imaging*, vol. 6357 of *Lecture Notes in Computer Science*, pp. 124–131, 2010.