Gland Segmentation in Hyperspectral Images using Unsupervised methods for Cancer detection in Colon

A B. Tech Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Bachelor of Technology

by

Vinay Chandragiri
(120101018)

under the guidance of

Dr. Amit Sethi & Dr. Saswata Shannigrahi

to the

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, ASSAM

CERTIFICATE

*This is to certify that the work contained in this thesis entitled "**Gland Segmentation in Hyperspectral Images using Unsupervised methods for Cancer detection in Colon** " is a bonafide work of **Vinay Chandragiri** (**Roll No. 120101018**), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati under my supervision and that it has not been submitted elsewhere for a degree.*

Guide: Dr. Amit Sethi,
Department of Electronics & Electrical Engineering,
Supervisor: Dr. Saswata Shannigrahi,
Department of Computer Science & Engineering,
Indian Institute of Technology Guwahati

Date: 12 November 2015

Motivation

The major motivation behind me working with Medical Image analysis is the one thing that's behind obtaining good results i.e contributing to a small community of people who are working for the betterment of lives of people by prevention and cure of deadly diseases. These Image Processing and Machine Learning techniques will be the key behind the success.

Abstract

Hyperspectral Imaging is the new modality in medical applications which is probably being used in Remote sensing applications. The image is generally of high dimension with spectral bands for a pixel. The main idea of the segmentation is to identify cancerous cells among the tissues. Here I am trying to address the problem of classifying cells by gland segmentation for cancer detection in the given colon tissue. The dimensionality problem has been tackled by Band Selection based on Independent Component Analysis.

Key Words: Hyperspectral Imaging, ICA, Clustering, Colon tissue

Acknowledgements

I would like to take this moment to express my deep and sincere gratitude to **Dr. Saswata Shannigrahi**, from the Department of Computer Science for permitting me to work with **Dr. Amit Sethi** from the Department of Electronics and Electrical Engineering.

I acknowledge with thanks for the kind of support, inspiration and constructive timely guidance which I received from **Dr. Amit**. I believe this experience will help me through out my career and also motivate me in pursing graduate studies in the field of Machine Learning precisely Deep Learning.

Vinay Chandragiri

Contents

1 Introduction **1**

1.1 Hyperspectral Images 1

1.2 Why Hyperspectral Imaging is used in Heath Care ? 3

1.3 Colon Cancer 3

1.4 How Hyperspectral Imaging helps in determining Colon Cancer ? 3

1.4.1 Challenges 4

1.5 Review of Prior works with HSI 4

2 Working with Hyperspectral Images **5**

2.1 Curse of Dimensionality of Hyperspectral Images 5

2.2 Feature Selection and Feature reduction 6

2.2.1 What are features ? 6

2.2.2 Which one would be better and why ? 6

2.2.3 Some techniques that I have explored in this process 7

3 Proposed Model **8**

3.1 Independent Component Analysis 8

3.1.1 When do we do ICA ? 8

3.1.2 Why ICA ? 8

3.1.3 Why not PCA ? 9

3.1.4 ICA overview 10

3.1.5 Fast ICA - The algorithm 10
3.1.6 Disadvantages of ICA 11
3.2 Band selection based on ICA 11
3.3 Results 12

4 Future Work **16**

4.1 References 16

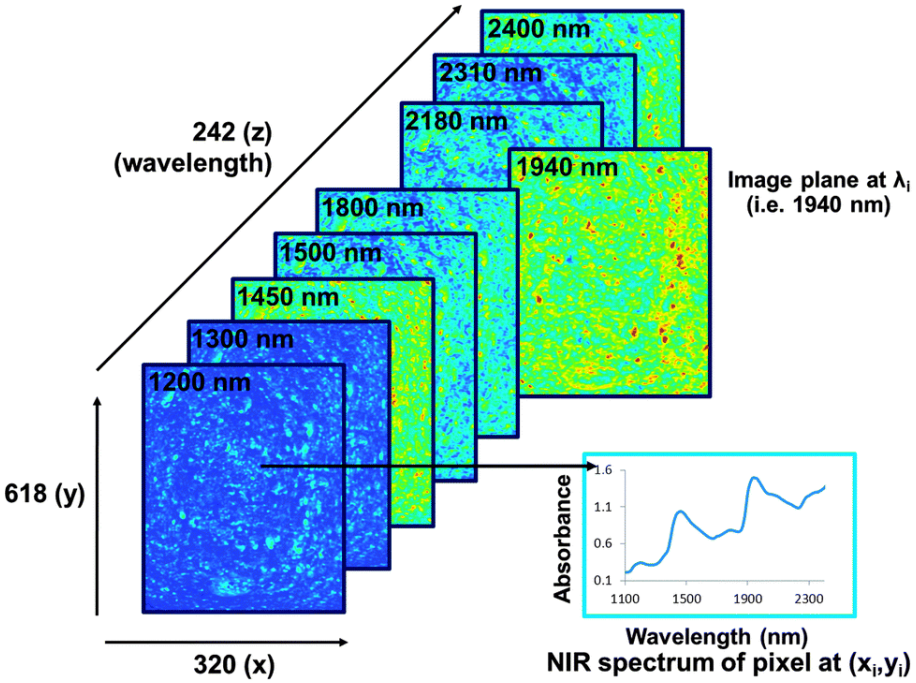
Chapter 1

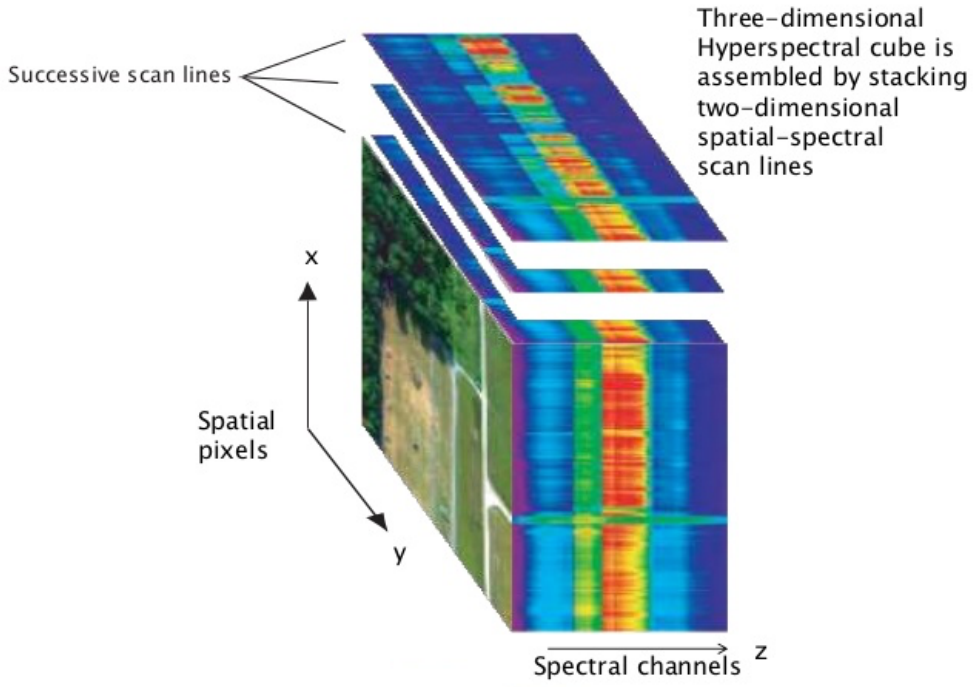
Introduction

1.1 Hyperspectral Images

Hyperspectral Imaging is a new and emerging technology in the field of Medical Imaging. It combines both digital imaging and spectroscopy. Instead of RGB values for each pixel, every pixel in the image contains a continous electromagnetic spectrum and is very useful in characterising the objects with immense accuracy and precision. Basically HSI is a stack of images(fig 1).

This new technology of Imaging is developed by NASA for Earth Imaging (Remote Sensing) and Space Observation. Hyperspectral Images are produced by instruments called Imaging Spectrometers.


2400 nm
2310 nm
2180 nm
1940 nm
1800 nm
1500 nm
1450 nm
1300 nm
1200 nm
242 (z)
(wavelength)
Image plane at λi
(i.e. 1940 nm)
618 (y)
320 (x)
Absorbance
1.6
1.1
0.6
0.1
1100
1500
1900
2300
Wavelength (nm)
NIR spectrum of pixel at (xi,yi)


Successive scan lines
Three-dimensional Hyperspectral cube is assembled by stacking two-dimensional spatial-spectral scan lines
x
Spatial pixels
y
Spectral channels
z

1.2 Why Hyperspectral Imaging is used in Heath Care ?

Exploiting the property of rich information in Hyperspectral Images (HSI), is a consistent development in the field of Medical Imaging and Health Care for various types of classification. It is found that this technology has important application in the field of Cancer detection. In medical image analysis of hyperspectral images provide a greater accuracy in determining the affected regions and reasons behind them.

HSI is capable of capturing both spectral information as well as spatial information in one shot. Hyperspectral Imaging provides a unique spectral signature, which can be used by processing techniques and discriminate materials.

1.3 Colon Cancer

The colon and rectum comprise the Large Intestine. Colon cancer is the third most common type of cancer after lung cancer and breast cancer. This is a disease in which normal cells in the lining of colon or rectum begin to change, grow without control and no longer die. It may cause bleeding which is painless and is not visible to eye.

1.4 How Hyperspectral Imaging helps in determining Colon Cancer ?

HSI helps in segmentation of glands and as well as classification of cells based on the properties of reflectance of different materials involved.

1.4.1 Challenges

In the **epithelial** lining of large intestine, intestinal glands are found. These glands contain different types of cells in that epithelium lining such as *goblet cells*, *enterocytes* etc. **Stroma** is a part of tissue that contributes to the connection and structures. Segmentation of these and grouping the similar types of cells together will help determine whether the tissue is benign or malignant if it is cancerous.

During Phase-I, I used statistical methods for classification (clustering) which are unsupervised. Supervised methods can yield better results with manual segmentation and using classifiers such as SVM's or Artificial Neural Networks.

1.5 Review of Prior works with HSI

Few recent advancements determine the extensive use of this technique in medical image analysis as the results are accurate. Prostate Cancer detection, Breast Cancer Detection and Skin Cancer Detection prevail among them.

Chapter 2

Working with Hyperspectral Images

2.1 Curse of Dimensionality of Hyperspectral Images

The term "Curse of Dimensionality" refers to the difficulties in processing the high dimensional data. In a single Hyperspectral Image *(say of dimension A* **x** *B* **x** *C)* , where A **x** B represents number of pixels and C represents the number of spectral. Considering the medical data used, A and B were in order of 500 - 600 and C in the order 220 - 240. So each pixel in that particular image is a C-dimensional vector and there are more than 2,00,00 pixels in one image.

This property of an HSI makes it tough to process the raw data for analysis for various purposes. This is the underlying reason leading to "dimensionality reduction" of Hyperspectral Images a very intensive research topic to work upon.

2.2 Feature Selection and Feature reduction

2.2.1 What are features ?

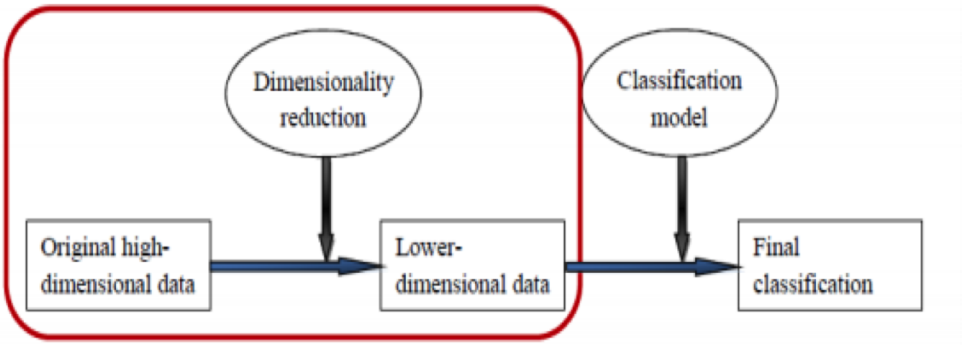
Features are attributes from raw data such that there value make an instance. These determinant values determine about the belonging of an instance to a particular class. We can classify features based on *Relevance*, *Irrelevence* and *Redunance*.

As the dimensionality problem has already been discussed earlier, to process the raw data, we can employ two different paradigms which are helpful in not using the redundant data for our processing. Redundancy exists when one feature takes place the role of another feature.

2.2.2 Which one would be better and why ?

Feature selection is a process by which we choose certain features *(a smaller subset)* from raw data using certain evaluation criteria. Different statistical methods such as clustering, minimizing **Mutual Information**, computing **KL divergence**, minimizing **Entropy** are explored. The main objectives of this process is to remove redundant data and irrelevant data.

Feature reduction (*also termed as Feature extraction)* is a process of mapping the High dimensional data to lower dimension such that the redundant and irrelevant data is minimized. *Principal component analysis* would be perfect example for this technique.



Optimally choosing a subset of features would be a better option instead of extracting new subsets of features as in the extraction will assure of loss of data and the process is irreversible. After doing some literature review about techniques that involve dimensionality reduction I turned myself over Feature selection.

2.2.3 Some techniques that I have explored in this process

* Principal Component Analysis
* Independent Component Analysis
* Multiple Discriminant Analysis

I have used ICA as a means to select most relevant bands from raw hyperspectral data instead of PCA and MDA.

Chapter 3

Proposed Model

3.1 Independent Component Analysis

3.1.1 When do we do ICA ?

* Raw data appears to be noisy

* When data is non gaussian i.e cannot be grouped via central limit theorem

* The sensor involved collects several source signals simultaneously

3.1.2 Why ICA ?

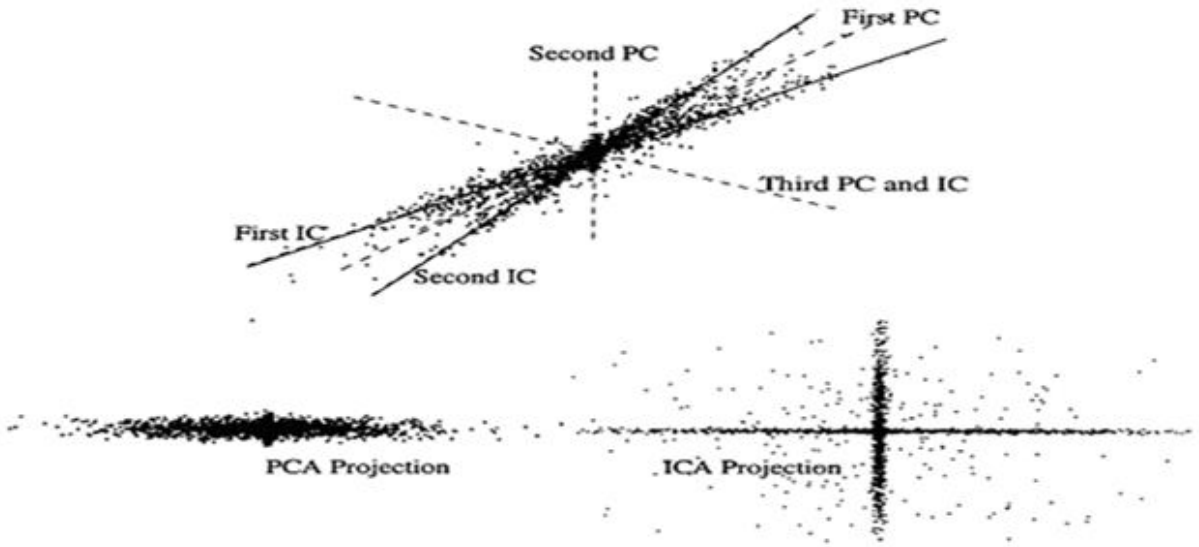
ICA minimizes the higher order statistics such or *Kurtosis* which will essentially minimize the mutual information in the output.

Note: This fact was explored and used to cluster the raw data directly with dimensionality reduction but was not successful as it turned out to be computationally expensive. Some of those results will be shown below.

3.1.3 Why not PCA ?

PCA minimizes the covariance of the raw data. PCA is most likely not suitable with Hyperspectral Images and these are highly correlated i.e the near by pixels are highly correlated.

The reason behind this is simple. With the uncorrelated variables in the data, the contributions by these in the lower dimension i.e to components will be almost equal in most of the cases. This does not happen with the variables which are correlated. We might skip the variables with most information with the process



3.1.4 ICA overview

Given a measurement X, assuming it as a linear combination of independent sources S and the mixing matrix A are to be determined such that $\mathbf{X = AS}$.

The independent sources are determined from the equation $\mathbf{S = (inv)A * X}$ i.e $\mathbf{S = WX}$ where W is termed as the mixing matrix.

3.1.5 Fast ICA - The algorithm

* *Input:* **X of (N x M)** represents N dimensional Sample. We can also fix the number of components we want to*(say C)*.

* *Output:* **W of (N x N)** the Un-mixing matrix.

for p **in** 1 to C:

$\mathbf{w_p} \leftarrow$ *Random vector of length N*

while $\mathbf{w_p}$ changes

$$\mathbf{w_p} \leftarrow \frac{1}{M}\mathbf{X}g(\mathbf{w_p}^T\mathbf{X}) - \frac{1}{M}g'(\mathbf{w_p}^T\mathbf{X})\mathbf{1w_p}$$

$$\mathbf{w_p} \leftarrow \mathbf{w_p} - \sum_{j=1}^{p-1} \mathbf{w_p}^T\mathbf{w_j}\mathbf{w_j}$$

$$\mathbf{w_p} \leftarrow \frac{\mathbf{w_p}}{\|\mathbf{w_p}\|}$$

Output: $\mathbf{W} = \begin{bmatrix} \mathbf{w_1} \\ \vdots \\ \mathbf{w_C} \end{bmatrix}$

Output: $\mathbf{S = WX}$

The key lies in determining the matrix **W** and a measure of non gaussianity.

3.1.6 Disadvantages of ICA

* We cannot determine the order of dominant components such as in PCA

* We cannot determine the variances of Independent components

3.2 Band selection based on ICA

This can be defined as the most crucial part of the work done in Phase - 1. Here, after obtaining the W from the fastICA, few points are to understood.

The absolute weight coefficients corresponding to each to each row are sorted and a band sequence is obtained because the weight matrix determines how a particular band contributes to each material in the hyperspectral image.

From the sequence, the bands with greater absolute weight coefficients will contribute more to the Independent transformations than other bands. This means that the bands with higher weights will have more spectral information than the other bands.

Thus from ICA, the bands are obtained and only those bands are selected from the original image and proceeded further as we ignored the irrelevant and redundant information is neglected.

These selected bands are called Independent bands and thus the unsupervised classification i.e clustering can be carried over with these selected bands. Basically, the new spectral image is constructed with these independent bands.

3.3 Results

The Segmented images are obtained after **K Means Clustering** the selected bands in the original hypersectral image into 3 & 4 clusters assuming there are 4 classes in the image. The results obtained for all the images in the dataset will be presented in the next chapter.

Other clustering algorithms have been explored such as **Mean Shift Clustering** and **Expectation Maximization** and will be given a try in Phase - 2

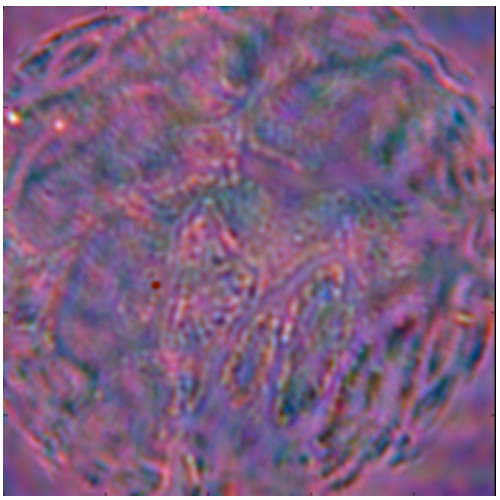


Fig. 3.1 Original Image

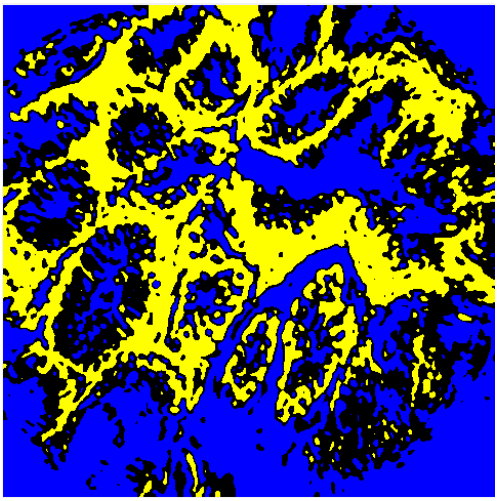


Fig. 3.2 Clustered Image (3 clusters) with 2 Spectral Bands 195 & 198 from all 226

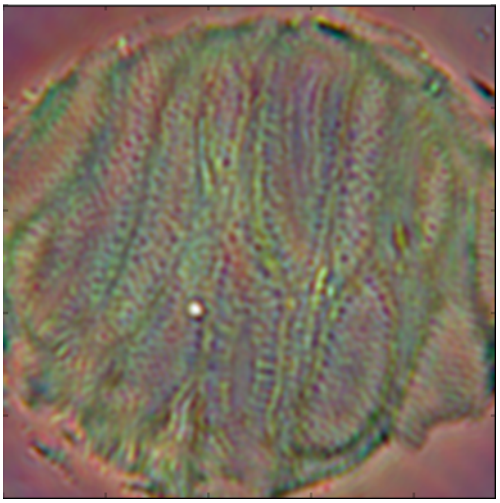


Fig. 3.3 Original Image

Fig. 3.4 Clustered Image (3 clusters) with 2 Spectral Bands 180 & 185 from all 226

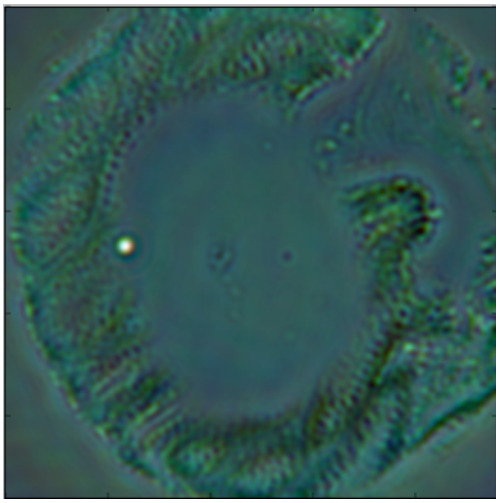


Fig. 3.5 Original Image

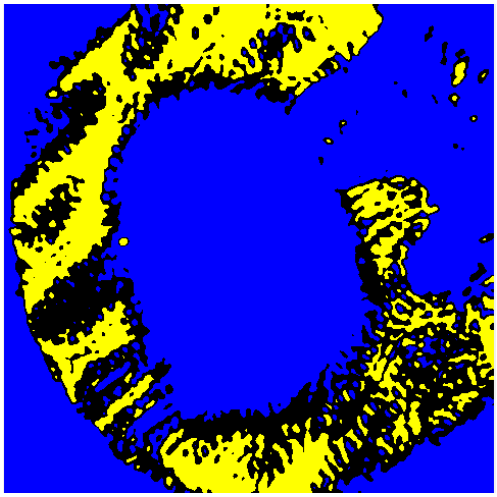


Fig. 3.6 Clustered Image (3 clusters) with 2 Spectral Bands 188 & 189 from all 226

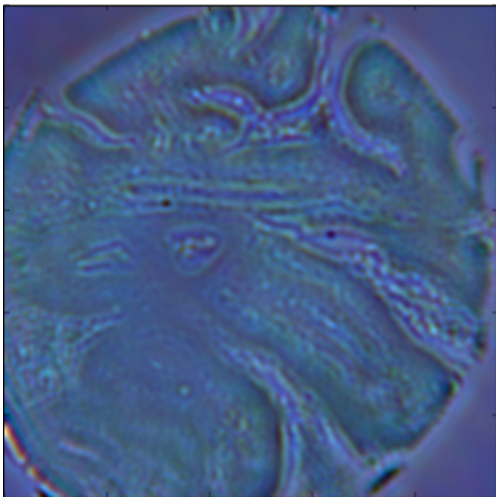


Fig. 3.7 Original Image

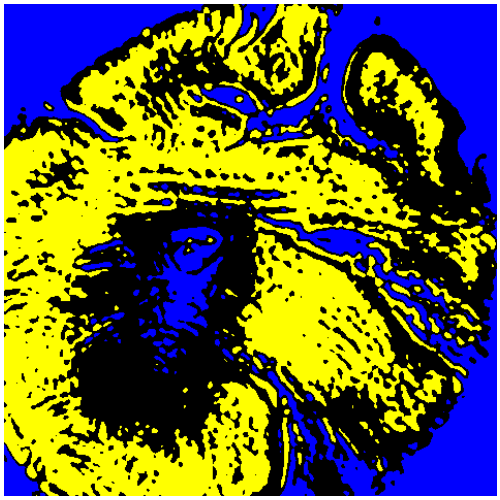


Fig. 3.8 Clustered Image with 2 Spectral Bands 181 & 182 from all 226

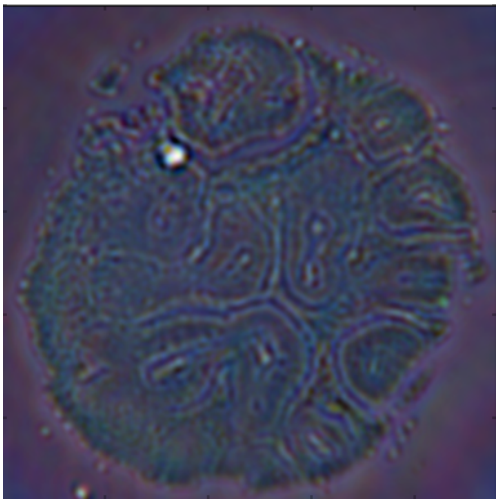


Fig. 3.9 Original Image

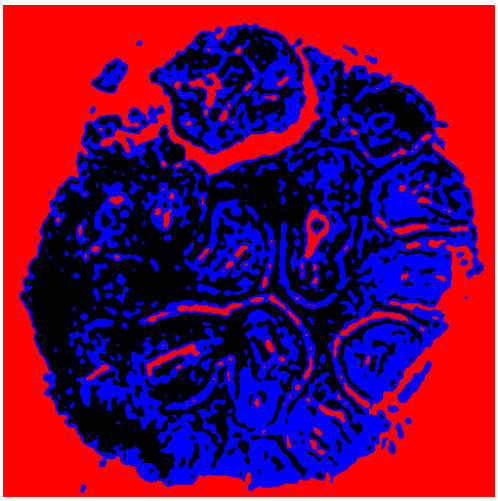


Fig. 3.10 Clustered Image (3 clusters) with 2 Spectral Bands 192 & 194 from all 226

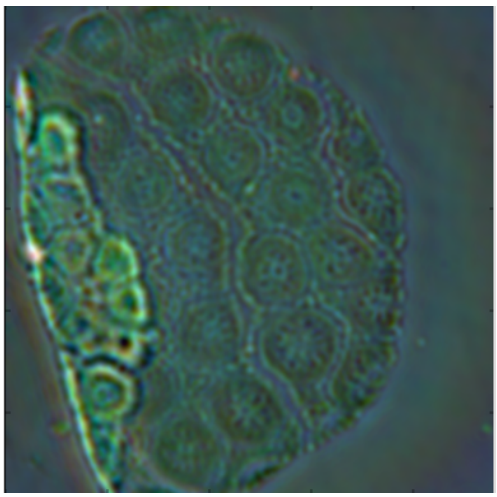


Fig. 3.11 Original Image

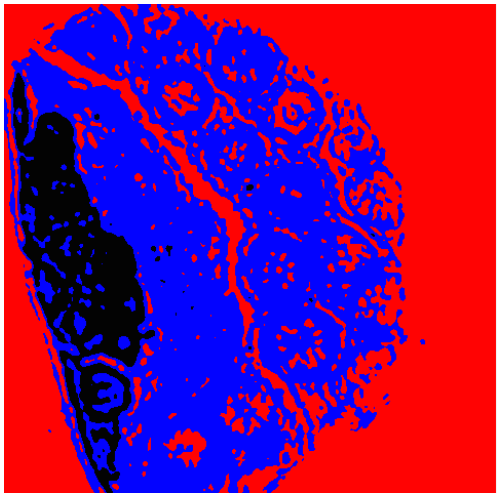


Fig. 3.12 Clustered Image with 2 Spectral Bands 200 & 216 from all 226

Chapter 4

Future Work

The Future work involves obtaining better structures of Glands. I am planning to try some supervised approaches with ANN or SVM or probably CNN to tackle the dimensionality and to obtain some promising results.

4.1 References

* Independent Component Analysis: Algorithms and Applications Aapo Hyvrinen and Erkki Oja Neural Networks Research Centre Helsinki University of Technology, Finland

* Epithelial-stromal interactions in colon cancer Fred T Bosman, The Netherlands

* Spatial Mutual Information Based Hyperspectral Band Selection for Classification

* www.wikipedia.org

* A Comparative Analysis of Dimension Reduction Algorithms on Hyperspectral Data Kate Burgers, Yohannes Fessehatsion, Jia Yin Seo, Sheida Rahmani

* A Novel Clustering-Based Feature Representation for the Classification of Hyperspectral Imagery Qikai Lu, Xin Huang and Liangpei Zhang