SEGMENTATION AND CLASSIFICATION OF CERVICAL CELL IMAGES

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Aslı Kale

January, 2010

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Selim Aksoy (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Pınar Duygulu Şahin

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Volkan Atalay

Approved for the Institute of Engineering and Science:

Prof. Dr. Mehmet B. Baray
Director of the Institute

ABSTRACT

SEGMENTATION AND CLASSIFICATION OF CERVICAL CELL IMAGES

Aslı Kale
M.S. in Computer Engineering
Supervisor: Assist. Prof. Dr. Selim Aksoy
January, 2010

Cervical cancer can be prevented if it is detected and treated early. Pap smear test is a manual screening procedure used to detect cervical cancer and precancerous changes in an uterine cervix. However, this procedure is costly and it may result in inaccurate diagnoses due to human error like intra- and inter-observer variability. Therefore, a computer-assisted screening system will be very beneficial to prevent cervical cancer if it increases the reliability of diagnoses.

In this thesis, we propose a computer-assisted diagnosis system which helps cyto-technicians by sorting cells in a Pap smear slide according to their abnormality degree. There are three main components of such a system. Firstly, cells along with their nuclei are located using a segmentation procedure on an image taken using a microscope. Then, features describing these segmented cells are extracted. Finally, the cells are sorted according to their abnormality degree based on the extracted features.

Different from the related studies that require images of a single cervical cell, we propose a non-parametric generic segmentation algorithm that can also handle images of overlapping cells. We use thresholding as the first phase to extract background regions for obtaining remaining cell regions. The second phase consists of segmenting the cell regions by a non-parametric hierarchical segmentation algorithm that uses the spectral and shape information as well as the gradient information. The last phase aims to partition the cell region into true structures of each nucleus and the whole cytoplasm area by classifying the final segments as nucleus or cytoplasm region. We evaluate our segmentation method both quantitatively and qualitatively using two data sets.

By proposing an unsupervised screening system, we aim to approach the problem in a different way when compared to the related studies that concentrate on classification. In order to rank the cells in a Pap slide, we first perform hierarchical clustering on 14 different cell features. The initial ordering of the cells is determined as the leaf ordering of the constructed hierarchical tree. Then, this initial ordering is improved by applying an optimal leaf ordering algorithm. The experiments with ground truth data show the effectiveness of the proposed approach under different experimental settings.

Keywords: Cytopathological image analysis, cell segmentation, hierarchical segmentation, ranking cells, computer-assisted diagnosis system, cervical cancer.

ÖZET

SERVİKS HÜCRE GÖRÜNTÜLERİNİN BÖLÜTLENMESİ VE SINIFLANDIRILMASI

Aslı Kale
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Yard. Doç. Dr. Selim Aksoy
Ocak, 2010

Serviks kanseri, erken teşhis ile tedavi edilerek önlenebilmektedir. Pap smear testi, rahim ağzında meydana gelen kanser ve kanser öncüsü değişiklikleri belirlemek üzere uygulanan manüel bir tarama yöntemidir. Ancak bu yöntem gözlemci tutarsızlığı ve her bir test için harcanması gereken çaba gibi dezavantajlar içermektedir. Bilgisayar destekli bir tarama sistemi, başarılı bir algoritma ile serviks kanserinin önlenmesinde yararlı olacaktır.

Bu tezde, verilen bir Pap test lamında yer alan hücreleri anormallik derecesine göre sıralayarak sitologlara yardımcı olacak bilgisayar destekli tanılayıcı bir sistem önerilmektedir. Böyle bir sitemi oluşturan üç temel bileşen vardır. İlk başta, hücreler ve çekirdekleri mikroskop kullanılarak elde edilen bir görüntü üzerinde bir bölütleme yöntemi yardımıyla tespit edilir. Sonra, bölütlenmiş olan hücreleri betimleyen özellikler çıkarılır. En sonunda, hücreler çıkarılan özellikler temel alınarak anormallik derecesine göre sıralanır.

Bir tek serviks hücresi içeren görüntüleri gerektiren ilgili çalışmalardan farklı olarak, örtüşen hücrelerin görüntülerini de işleyebilen parametrik olmayan genel bir bölütleme algoritması önerilmektedir. İlk aşama olarak, arka plan alanlarını çıkararak geriye kalan hücre alanlarını elde etmek amacıyla eşikleme yöntemi kullanılmıştır. İkinci aşama, spektral, şekil ve gradyan bilgisinden faydalanan parametrik olmayan hiyerarşik bir bölütleme yöntemi ile hücre alanlarının bölütlenmesinden oluşmaktadır. Son aşama, elde edilen bölütleri çekirdek ya da sitoplazma olarak sınıflandırmak suretiyle hücre alanını her bir çekirdeğe ait doğru yapılara ve bütün sitoplazma alanına ayırmayı amaçlamaktadır. Önerilen bölütleme yöntemi iki farklı veri kümesi kullanılarak nicel ve nitel olarak değerlendirilmiştir.

Öğreticisiz bir tarama sistemi önerilerek, sınıflandırma üzerine yoğunlaşmış ilgili çalısmalara göre probleme farklı bir yönden yaklaşmak amaçlanmıştır. Bir Pap lamında yer alan hücreleri sıralamak için, ilk önce, hücrelerden çıkarılan 14 farklı özelliğe göre hiyerarşik kümeleme uygulanmıştır. Hücrelerin ilk sıralaması oluşturulan hiyerarşik ağacın yaprak sıralaması olarak belirlenmiştir. Sonra, bu ilk sıralama bir en iyi yaprak sıralama algoritması ile iyileştirilmiştir. Referans veri kullanılarak yapılan deneyler önerilen yaklaşımın etkinliğini farklı deneysel ayarlar altında göstermektedir.

Anahtar sözcükler: Sitopatolojik görüntü analizi, hücre bölütlemesi, hiyerarşik bölütleme, hücrelerin sıralanması, bilgisayar destekli tanılayıcı sistem, serviks kanseri.

Acknowledgement

I would like to thank my advisor Assist. Prof. Dr. Selim Aksoy for his great supervision and invaluable helps throughout this study. It has always been a pleasure to work with him and get benefit from his vision and knowledge in every step of my research.

I would also thank Assist. Prof. Dr. Pınar Duygulu Şahin and Prof. Dr. Volkan Atalay for kindly agreeing to be in my thesis committee. I thank to Dr. Sevgen Önder for his consultancy on medical knowledge and to Hacettepe University, Department of Pathology, for providing us the dataset.

I would like to express my gratitude to my family for their endless support and love. I would also like to thank Murat for his always being with me.

I would like to thank Gökhan for his guidance and comments. I am grateful to my friends Daniya, Sare, Fırat, Bahadır, Onur, Selen, Çağlar, Nazlı and other RETINA members for their nice friendship.

I also express my pleasure to TÜBİTAK (The Scientific and Technological Research Council of Turkey) for supporting me financially. This work was also supported in part by the TÜBİTAK CAREER Grant 104E074.

Contents

1 Introduction **1**

1.1 Overview 1

1.2 Problem Definition 2

1.3 Data Set 8

1.3.1 Herlev Data Set 8

1.3.2 Hacettepe Data 9

1.4 Summary of Contributions 10

1.5 Organization of the Thesis 11

2 Literature Review **12**

2.1 Segmentation of Cervical Cells 12

2.2 Classification of Cervical Cells 15

3 Segmentation of Cervical Cells **17**

3.1 Background Extraction 18

3.2 Nucleus and Cytoplasm Segmentation 29

3.2.1 Hierarchical Region Extraction 30
3.2.2 Region Selection 46
3.3 Nucleus and Cytoplasm Classification 51
3.3.1 Bayesian Classifier for Non-parametric Densities 52
3.3.2 Bayesian Classifier for Normal Densities 55
3.3.3 Decision Tree Classifier 57
3.3.4 Support Vector Classifier 58
3.3.5 Combined Classifiers 59

4 Classification of Cervical Cells 62

4.1 Feature Extraction 62
4.2 Ranking of Cervical Cells 65

5 Experiments and Results 72

5.1 Segmentation of Cervical Cells 72
5.1.1 Background Extraction 73
5.1.2 Nucleus and Cytoplasm Segmentation 77
5.1.3 Nucleus and Cytoplasm Classification 89
5.2 Ranking of Cervical Cells 89
5.2.1 Rank-Order Correlation Coefficients 90
5.2.2 Kappa Coefficients 91
5.2.3 Experimental Results 93

6 Conclusion and Future Work **99**

List of Figures

1.1 A cervical cell stained with hematoxylin-and-eosin. 3

1.2 Pap smear images stained with hematoxylin-and-eosin taken at different magnifications (a) 20x (b) 40x (c) 100x. 4

1.3 The block diagram of the proposed system. 7

1.4 An example cell image and its segmentation result from the Herlev data set. 9

3.1 An example Pap smear image with its characteristic problems such as inhomogeneous staining, overlapping cells. 18

3.2 A Pap smear image (a) in RGB color space (b) L channel of the transformed image in CIE Lab color space (c) the histogram of the L channel. 20

3.3 Use of black top-hat transform for mitigating inhomogeneous illumination (a) L channel of the image (b) closing with a large structuring element (c) the illumination-corrected L channel by the black top-hat transform (d) the histogram of the illumination-corrected L channel. 21

3.4 (a) The histogram taken from Figure 3.3 (b) its corresponding criterion function. 23

3.5 (a) The Pap smear image taken from Figure 3.2 (b) regions found by thresholding at $T = 0.03$ (c) cell regions after eliminating small areas (d) boundaries of cell regions. 25

3.6 Illumination-corrected channels of RGB color space, their histograms and segmentation results (a) R channel (b) G channel (c) B channel. 26

3.7 Illumination-corrected channels of CIE XYZ color space, their histograms and segmentation results (a) X channel (b) Y channel (c) Z channel. 27

3.8 Channels of CIE Lab color space, their histograms and segmentation results (a) illumination-corrected L channel (b) a channel (c) b channel. 28

3.9 Candidate segments obtained by morphological profiles of closing by reconstruction (a) at SE size 1 (b) at SE size 4 (c) at SE size 7 (d) at SE size 10 (e) at SE size 13 (f) at SE size 15. 32

3.10 Candidate segments obtained by morphological profiles of opening by reconstruction (a) at SE size 1 (b) at SE size 4 (c) at SE size 7 (d) at SE size 10 (e) at SE size 13 (f) at SE size 15. 33

3.11 Hierarchical watershed transform based on dynamics. (Image taken from [13].) 34

3.12 (a) One-dimensional signal f (b) marker f_m (c) point-wise minimum between $f + 1$ and f_m (d) reconstruction by erosion of (c) from f_m. 36

3.13 Candidate segments obtained by multi-scale watershed segmentation based on dynamic. (a) at scale 0 (b) at scale 1 (c) at scale 13 (d) at scale 14 (e) at scale 26 (f) at scale 27. 37

3.14 (a) One dimensional signal (blue) (b) h-minima transformations (red) for $h = 1$ (c) $h = 2$ (d) $h = 3$. 39

3.15 (a) Cell image (b) its gradient (c) the minima at the raw gradient (d) the minima at the h-minima of the gradient for $h = 17$. 39

3.16 One dimensional signal (blue) and watersheds (black) of h-minima transformations (red) (a) at scale 0 (b) at scale 1 (c) at scale 2 (d) at scale 3 (e) at scale 4 (f) at scale 5 (g) at scale 6 (h) at scale 7. . 40

3.17 Candidate segments obtained by multi-scale watershed segmentation based on h-minima transform. (a) at scale 2 (b) at scale 3 (c) at scale 6 (d) at scale 7 (e) at scale 12 (f) at scale 13. 41

3.18 One-dimensional signal (green) and watersheds (black) of transformed signal (red) by minima imposition at scale (a) 0 (b) 1 (c) 2 (d) 3 (e) 4 (f) 5 (g) 6 (h) 7. 42

3.19 One-dimensional signal (green) and watersheds (black) of transformed signal (red) by h-minima at scale (a) 0 (b) 1 (c) 2 (d) 3 (e) 4 (f) 5 (g) 6 (h) 7. 43

3.20 (a) Watershed partition of one-dimensional signal at scale 0 (b) partition of the second method at scale 1 (c) partition of the third method at scale 1 (d) each partition of scale 1 is adjusted. 44

3.21 An example tree. Node i_j is a segment of the partition at scale i. j denotes the sequence number of the node from left to right in level i. 45

3.22 An example run of the bottom-up algorithm for the example tree given in Figure 3.21. Starting from level 1, the nodes having a measure greater than all of its descendants are colored with blue in each step. 50

3.23 An example run of the top-down algorithm for the example tree given in Figure 3.21. Starting from the root level, the nodes marked in the first pass while none of its ancestors is marked are marked as *selected*. The green nodes are the final most meaningful segments. 50

3.24 The segmentation result of the example cell image. 51

3.25 Class conditional probability density functions for the feature components (a) size (b) mean intensity (c) eccentricity (d) homogeneity. 54

3.26 Likelihood ratio threshold versus accurate classification rate for nucleus (red) and cytoplasm (blue) segments given feature combinations (a) mean intensity (b) mean intensity and eccentricity (c) mean intensity, eccentricity and size (d) mean intensity, eccentricity, size and homogeneity. 56

3.27 (a) The segmentation result (b) the classification result of the example cell image. 61

4.1 (a) The example cell image, and (b) the corresponding segmentation and classification result. 63

4.2 An example nucleus region (green) surrounded by cytoplasm region (blue). Longest diameter line L and shortest diameter lines S_1 and S_2 are shown. 64

4.3 The binary tree resulting from hierarchical clustering of 30 cells randomly selected from the Herlev data. We select 5 cells from each class in order of normal superficial (1−5), normal intermediate $(6-10)$, mild dysplasia $(11-15)$, moderate dysplasia $(16-20)$, severe dysplasia $(21-25)$, and carcinoma in situ $(26-30)$. 67

4.4 (a) An example binary tree T and (b) a linear leaf ordering consistent with T obtained by flipping the node marked by red ellipse. 68

4.5 The initial ordering of the cells determined as the linear ordering of the leaves of the tree in Figure 4.3. (The cell images are resized to have the same width and height so their relative size is not proper.) 70

4.6 The final ordering of the cells obtained by applying the optimal leaf ordering algorithm to the initial ordering in Figure 4.5. (The cell images are resized to have the same width and height so their relative size is not proper.) 71

5.1 (a) The histogram of the L channel (b) the histogram of the illumination-corrected L channel (c) the boundaries of the segmented cell regions are colored as red. 74

5.2 (a) The histogram of the L channel (b) the histogram of the illumination-corrected L channel (c) the boundaries of the segmented cell regions are colored as red. 75

5.3 (a) The histogram of the L channel (b) the histogram of the illumination-corrected L channel (c) the boundaries of the segmented cell regions are colored as red. 76

5.4 The ZSIs for the images of the classes (a) Intermediate squamous (b) Superficial squamous (c) Columnar 78

5.5 The ZSIs for the images of the classes (a) Mild dysplasia (b) Moderate dysplasia (c) Severe dysplasia (d) Carcinoma in situ 79

5.6 Segmentation results for example images from Hacettepe data. . . 81

5.7 Segmentation results for example images from Hacettepe data. . . 82

5.8 Segmentation results for example images from Hacettepe data. . . 83

5.9 Segmentation results for example images from Hacettepe data. . . 84

5.10 Segmentation results for example images from Hacettepe data. . . 85

5.11 Problematic segmentation results for example images from Hacettepe data. 87

5.12 Segmentation result for an example image consisting of many overlapping noisy cells. 88

5.13 The first ordering of the cells where $r_s = 0.895$, $D = 792$, $\kappa = 0.466$ and $\kappa_w = 0.614$. (The cell images are resized to have the same width and height so their relative size is not proper.) 96

5.14 The second ordering of the cells where $r_s = 0.771$, $D = 1728$, $\kappa = 0.266$ and $\kappa_w = 0.417$. (The cell images are resized to have the same width and height so their relative size is not proper.) . . 97

5.15 The third ordering of the cells where $r_s = 0.038$, $D = 7272$, $\kappa = -0.100$ and $\kappa_w = -0.054$. (The cell images are resized to have the same width and height so their relative size is not proper.) 98

List of Tables

1.1 Normal cervical cells and their characteristics. 5

1.2 Abnormal cervical cells and their characteristics. 6

1.3 Distribution of the Herlev data among 7 classes. 8

3.1 Classification performances of different classifiers. The number of misclassified nucleus (N), cytoplasm (C) and total (T) regions in the testing data set are given for each classifier based on both original and normalized features. 60

5.1 The ZSI means and standard deviations of each class for the ground truth compared to our segmentation. 80

5.2 An example ranking scenario. 90

5.3 The experimental results for the ranking of cervical cells obtained for different settings. 95

Chapter 1

Introduction

1.1 Overview

Worldwide, cervical cancer has a significant impact, with nearly $500,000$ new cases and nearly $250,000$ deaths reported annually [17]. Cervical cancer develops over a long period of time. It usually takes many years for cervical cancer to progress from a benign to a life-threatening stage. Moreover, symptoms of this cancer may be absent until it is in its advanced stages and at these late stages, the cancer is usually unresponsive to treatment [44].

Pap smear test is used to detect cervical cancer and precancerous changes in an uterine cervix. Precancerous changes in cervical cells are called dysplasia. A cervical cell turns into a precancerous cell when it does not divide as it should due to some change in the genetic information of the cell [26]. Cervical cancer can be prevented through screening at-risk women and treating women with precancerous and cancerous lesions. Incidence and mortality rates have decreased steadily over the past five decades, largely due to the widespread use of the Pap smear [17].

Pap smear test is a manual screening procedure that requires well skilled cytotechnicians. This procedure is costly and it may result in inaccurate diagnoses due

to human error like intra- and inter-observer variability. Therefore, a computer-assisted screening system will be very beneficial to prevent cervical cancer if it increases the reliability of diagnoses.

In this thesis, we propose a computer-assisted diagnosis system which helps cyto-technicians by sorting cells in a Pap smear slide according to their dysplasia degree. There are three main components of such a system. Firstly, cells along with their nuclei are located using a segmentation procedure on an image taken using a microscope. Then, features describing these segmented cells are extracted. Finally, the cells are sorted from normal to abnormal based on the extracted features.

1.2 Problem Definition

Pap smear test is a medical procedure to detect precancerous or cancerous cells in the uterine cervix. In this test, a specimen is taken from the uterine cervix and smeared onto a thin rectangular glass slide using a special cyto-brush. Then, the cells on the slide are colored by generally using hematoxylin-and-eosin stain to make their examination easier. By this way, cyto-technicians can diagnose premalignant cell changes under the microscope before they progress to a cancer. A stained cell image which contains a nucleus surrounded by cytoplasm on a background is shown in Figure 1.1. Example Pap smear images taken at different magnifications are given in Figure 1.2 and it can be seen that details of the cells become more apparent as the image magnification increases.

A single Pap smear slide may contain hundreds of thousands of cells and cyto-technicians examine these cells under the microscope to determine premalignant cell changes based on the cell characteristics like size, color, shape and texture of nucleus and cytoplasm.

A cervical cell can be mainly diagnosed as normal or abnormal. Table 1.1 shows characteristics of normal cervical cells located at separate areas of the

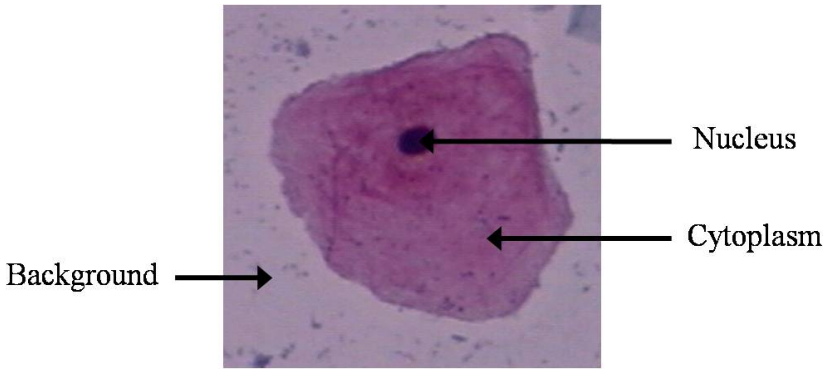
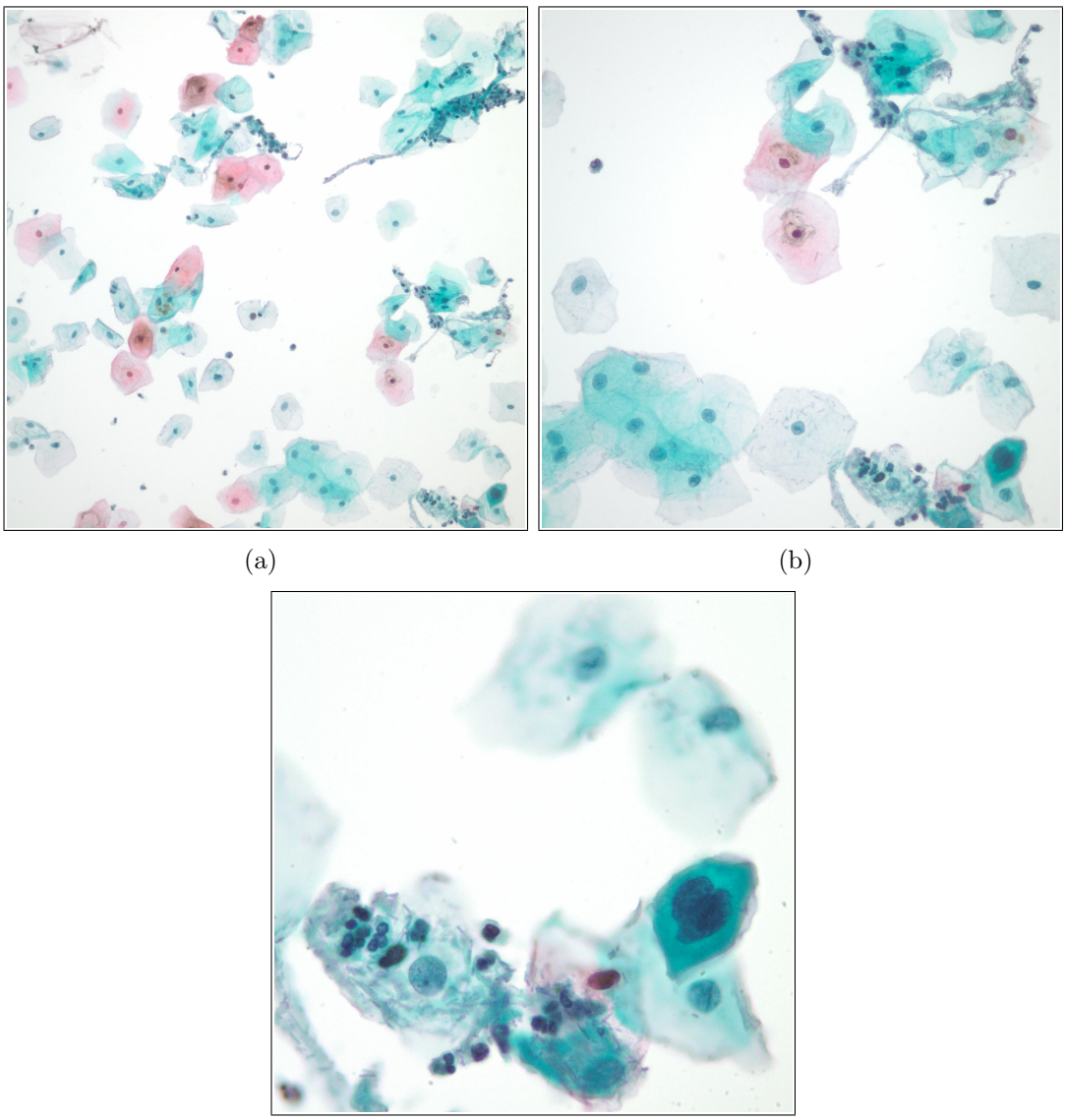


Figure 1.1: A cervical cell stained with hematoxylin-and-eosin.

cervix, namely squamous and columnar area. Superficial and intermediate squamous cells lie on different layers of the squamous area. When intermediate squamous cells mature, they move to superficial layer and become superficial squamous cells. While moving through the layers, the cytoplasm becomes bigger and the nucleus becomes smaller. Characteristics for columnar cells are a column-like shape with an oblong cytoplasm and a large nucleus located at one end.

Table 1.2 shows example abnormal cervical cells of different categories and their characteristics. Categories of abnormal cells describe the risk that cells turn into malignant cancer cells. For example, mild dysplastic cells have lower risk of becoming malignant cancer cells than severe dysplastic cells. Mildly dysplastic cells have enlarged and bright nuclei. Moderately dysplastic cells have even larger and darker nuclei. The nuclei may start to deteriorate. Severe dysplastic cells have large, dark, and deformed nuclei and their cytoplasm is relatively dark and small. Characteristics of cells in carcinoma in situ are similar to the ones in severe dysplasia. As can be seen from Table 1.1 and Table 1.2, precancers and cancers are associated with a variety of morphological and architectural changes, including size, texture, and shape of nucleus and cytoplasm along with the increasing ratio of nucleus and cytoplasm area.

In this thesis, we propose to rank cervical cells in a Pap smear slide according to their abnormality degree. The block diagram of our proposed system is shown in Figure 1.3. Given an input Pap test image, we first extract the background region in order to obtain the remaining cell regions. Then, we apply our



(a) (b)

(c)

Figure 1.2: Pap smear images stained with hematoxylin-and-eosin taken at different magnifications (a) 20x (b) 40x (c) 100x.

Table 1.1: Normal cervical cells and their characteristics.

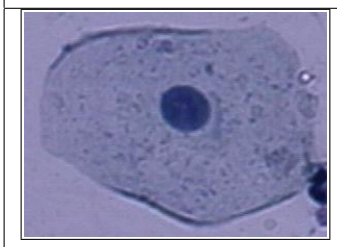
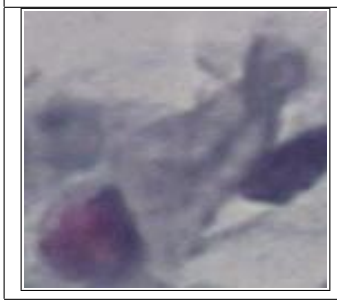
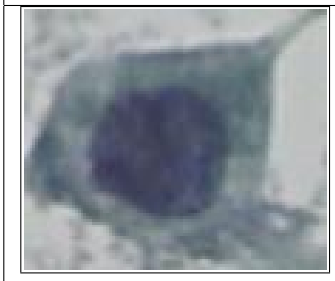
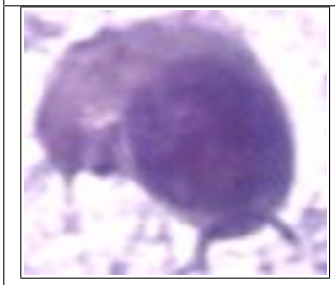
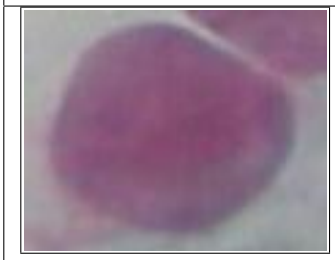
Normal cells	
	**Superficial squamous** • Shape flat/oval • Nucleus very small • Nucleus/cytoplasm ratio very small
	**Intermediate squamous** • Shape round • Nucleus large • Nucleus/cytoplasm ratio small
	**Columnar** • Shape column-like • Nucleus large • Nucleus/cytoplasm ratio medium

Table 1.2: Abnormal cervical cells and their characteristics.

Abnormal cells	
	**Mild dysplasia** • Nucleus light/large • Nucleus/cytoplasm ratio medium
	**Moderate dysplasia** • Nucleus large/dark • Cytoplasm dark • Nucleus/cytoplasm ratio large
	**Severe dysplasia** • Nucleus large/dark/deform • Cytoplasm dark • Nucleus/cytoplasm ratio very large
	**Carcinoma in situ** • Nucleus large/dark/deform • Nucleus/cytoplasm ratio very large

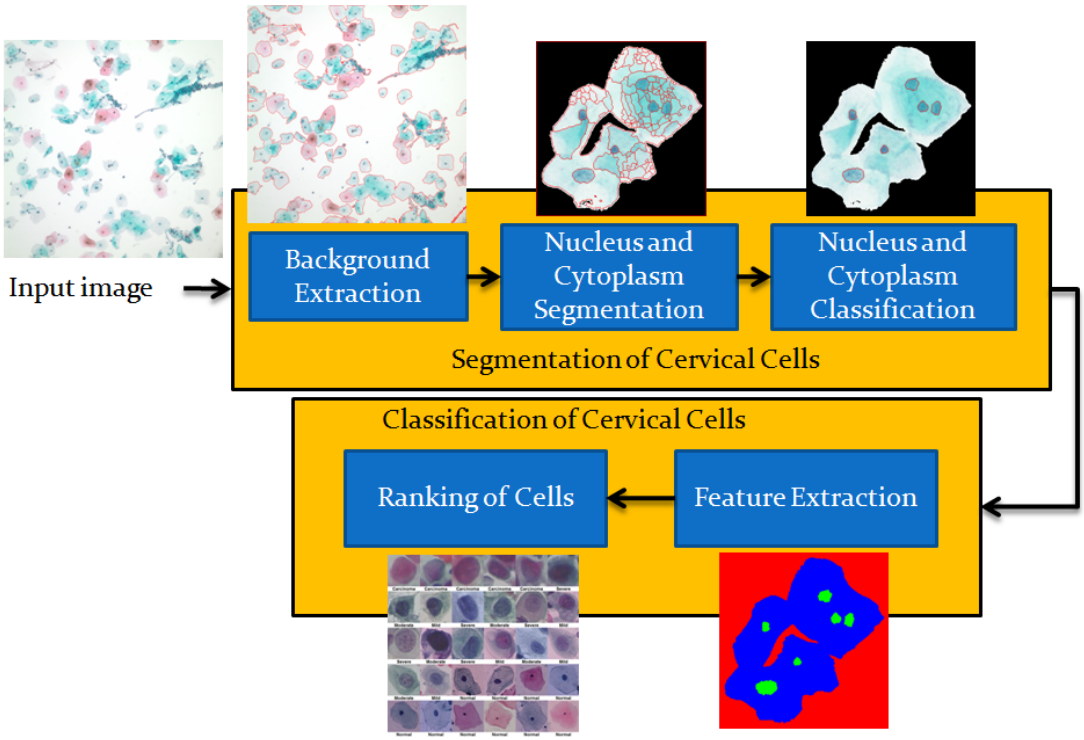


Figure 1.3: The block diagram of the proposed system.

non-parametric hierarchical segmentation algorithm to each cell region. In the segmentation step, our primary goal is to obtain a corresponding region for the true structure of each nucleus. After that we classify the segments as nucleus or cytoplasm area using a sum combination of a Bayesian classifier, a support vector classifier and a decision tree classifier. At this point, the whole cytoplasm area is determined as the union of all cytoplasm segments and the nucleus segments constitute true structures of each nucleus. After dividing each cell region into true structures of each nucleus and the remaining whole cytoplasm area, we extract 14 different features from each cell which is denoted by its nucleus region. In order to rank the cells, we first perform hierarchical clustering on the extracted cell features where each cell is a leaf in the cluster hierarchy. The linear leaf ordering of the constructed tree is considered as a ranking of the cells, and this ranking is further improved using the fast optimal leaf ordering algorithm proposed in [6].

Table 1.3: Distribution of the Herlev data among 7 classes.

Normal cells	
Superficial squamous	74 cells
Intermediate squamous	70 cells
Columnar	98 cells
Abnormal cells	
Mild dysplasia	182 cells
Moderate dysplasia	146 cells
Severe dysplasia	197 cells
Carcinoma in situ	150 cells

1.3 Data Set

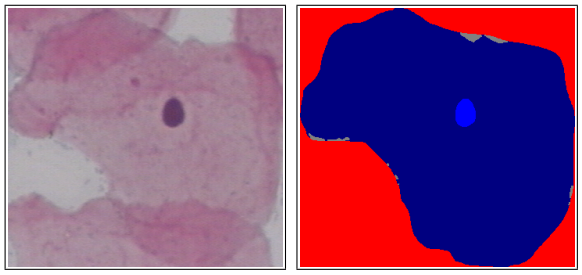
The methodologies presented in this thesis are illustrated using two different data sets.

1.3.1 Herlev Data Set

The Herlev data set consists of 917 images of single Pap smear cells [23]. It was developed by the Department of Pathology at Herlev University Hospital and the Department of Automation at Technical University of Denmark to provide benchmark data for comparing classification methods.

The data was collected by cyto-technicians using a microscope connected to a digital camera. Each cell image was taken with a magnification of $0.201 \mu m/pixel$. Cyto-technicians and doctors manually classified each cell into one of the 7 classes described in Table 1.1 and 1.2. Each cell was examined by two cyto-technicians, and difficult samples were also checked by a doctor. In case of disagreement, the sample was discarded. Thus, the data set contains diagnoses that are as certain as possible. Table 1.3 shows the distribution of the data set among 7 classes.

All images in the data set were segmented into background, cytoplasm, and nucleus regions using the CHAMP software. CHAMP is a commercial medical



(a) (b)

Figure 1.4: An example cell image and its segmentation result from the Herlev data set.

image analysis system that uses a patented object recognition method [23]. The segmentation results were examined by the cyto-technicians to see if was is necessary and possible to make a correction and the images whose segmentation failed were removed from the data set. Figure 1.4 shows an example cell image and its segmentation result. After segmentation, 20 different features describing cell characteristics like size, area, shape and brightness of both nucleus and cytoplasm were extracted from each cell.

1.3.2 Hacettepe Data

The Hacettepe data set was prepared by the Department of Pathology at Hacettepe University Hospital and the Department of Computer Engineering at Bilkent University. The data was collected by Dr. Sevgen Onder from Hacettepe University using a microscope connected to a digital camera.

It consists of 198 Pap test images taken at different magnifications. There are 82 images taken at 20x magnification, 84 images taken at 40x magnification and 32 images taken at 100x magnification. The data was collected from the Pap test slides of 18 different patients. The size of each image is 2048x2048 pixels. Example images taken at different magnifications are shown in Figure 1.2. Only the images at 20x magnification are used in this thesis.

1.4 Summary of Contributions

In this thesis, our goal is to rank the cells in a Pap smear slide according to their abnormality degree. In this way, the cells that are ranked as more normal than a selected cell that is manually identified as normal can be skipped by cytotechnicians or the cells can be investigated beginning from the end of the rank list that the most abnormal cells are found.

The first and the most crucial step of the proposed system is the accurate segmentation of cells along with their nucleus and cytoplasm. The nature of the Pap test images consisting of many overlapping cells makes the related studies requiring images of a single cell impractical. Thus, we propose a three-phase generic segmentation approach where thresholding is used as the first phase to extract the background regions for obtaining the remaining cell regions. The second phase consists of segmenting the cell regions by a non-parametric hierarchical segmentation algorithm. The last phase aims to partition the cell region into true structures of each nucleus and the whole cytoplasm area by classifying the final segments as nucleus or cytoplasm region.

Our segmentation method follows the general framework of the method developed by Akçay and Aksoy [1]. The main difference and advantage of our approach stems from being a non-parametric hierarchical segmentation algorithm that uses the spectral and shape information as well as the gradient information. Instead of using morphological opening and closing by reconstruction operations in [1], we extract the candidate regions by applying watershed segmentation to h-minima transforms of the image gradient for increasing values of h. Then, we similarly construct a hierarchical tree from the extracted regions and select the most meaningful regions in that tree by optimizing a measure. However, the measure we use is different from the one used in [1] such that our measure consists of two factors as the spectral homogeneity, which is calculated in a different way, and the circularity. We evaluate our segmentation approach both quantitatively and qualitatively on two data sets one of which was developed by the Department of Pathology at Hacettepe University and the Department of Computer Engineering at Bilkent University. The Hacettepe data consist of more realistic examples

compared to the images of other data sets used in the literature.

Unlike existing studies that aim to classify individual cells using supervised classifiers, we approach the classification process using an unsupervised ranking procedure. Apart from two FDA approved commercial devices [12], as far as we know, there is no similar work in the literature. In order to rank cells, we employ the fast optimal leaf ordering algorithm [6] used in the biological literature to explore related genes that share a common function.

1.5 Organization of the Thesis

The rest of the thesis is organized as follows. In Chapter 2, we give summary about the related works. In Chapter 3, we describe our segmentation method in detail. In Chapter 4, we first present the extracted cell features and then describe the method we use to rank cervical cells. In Chapter 5, we present our experimental results. Finally, conclusions and future research directions are given in Chapter 6.

Chapter 2

Literature Review

In this section, we first discuss some of the previous works on segmentation of cells. We then give the review of the studies related to classification and ranking of cervical cells.

2.1 Segmentation of Cervical Cells

Below, we present a survey of the methods related to segmentation of various types of cells.

Bamford and Lovell [5] propose a method for segmenting cervical cells along with their nucleus. In their approach, conventional Pap test slides are first scanned at a low magnification to find the locations of the cells using the algorithm presented in [4]. Once the cells are located for further examination, they are reviewed at a higher magnification. Since nuclei are darker than surrounding cytoplasm, they first find a point within the nucleus by using one of two techniques, namely converging squares [27] and simple thresholding. Then, they construct a search space by two concentric circles, one lying within the nucleus and the other lying outside. After discretizing the search space, they apply the Viterbi algorithm to find the global minimum contour of the nucleus according

to some cost function. This method fails when there is a very large image gradient near nucleus border or the regularization parameter for finding the global minimum contour is selected inappropriately.

Yang-Mao et al. [44] propose a detector for nucleus and cytoplasm contour of a cervical cell. They first enhance the edges at the borders of nucleus and cytoplasm by applying a series of operations, namely the trim-mean filter, the bigroup enhancer and the mean vector difference enhancer defined in [44]. Then, the contours of nucleus and cytoplasm are obtained based on thresholding approach. This method requires the images of a single cervical cell and it is not suitable for our problem involving overlapping cells.

Wu et al. [40, 42] introduce a parametric optimal segmentation approach which requires prior knowledge about the nucleus characteristics like the shape, size and intensity of the nucleus relative to its surrounding area. They model a cell image as an elliptical nucleus region which has two level intensities inside and outside the region. Then, a cost for a parameter set by taking difference of the original image and proposed image of the model is calculated. They obtain the final model by finding the parameters of the model that lead globally the minimum cost. By thresholding the final model, the corresponding segmentation result is produced. This method is not suitable for overlapping cells because its computational complexity becomes high.

Walker et al. [38] use a series of automated fast morphological transforms with octagonal structuring elements to segment each nucleus from its cytoplasm. They first perform global thresholding on the cell image to obtain the incomplete segmentation of the nucleus in binary form. Cytoplasmic backgrounds are removed by morphological closing using a structuring element smaller than the smallest nucleus and the nucleus areas are corrected by morphological opening using the same size for the structuring element. However, this method suffers from the global thresholding and it can be improved using local thresholding methods.

In order to segment immunohistochemically stained images, Shah [32] proposes a two phase approach that combines the low level operations like clustering with the higher level operations such as classification. The first approximation of

the cell locations are calculated by an unsupervised clustering approach coupled with cluster merging based on a fitness function. Then, a joint segmentation classification method incorporating ellipse as a shape prior is used in the second phase to obtain the final cell locations. This method is best suited to the Pap test images taken at lower magnifications compared to our images.

Dagher and Tom [11] introduce a new segmentation technique by combining the watershed algorithm and the active contour model. They apply watershed transform on the image gradient after filtering small noisy regions. Then, the contours of the obtained segments are used for the initialization of the snake model. Once the snake captures one object, the image is relabeled so that the next snake will not be a watershed contour inside the captured object. Note that the active contour models [20, 43] may suffer from the initialization, parameter selection and small capture range problems. They use the Balloon snake model [10] to solve the problem of small capture range and a new parameter optimization approach is also proposed. This method is applied to the images in which each cell appears as a whole homogeneous region rather than union of nucleus and cytoplasm area. We can use this approach to segment nucleus of cervical cells but finding initial segments related to nucleus regions then becomes a problem.

A system for grading hepatocellular carcinoma biopsy images is proposed by Huang and Lai [16] where the images are classified based on the features extracted after the segmentation of nucleus regions. In their segmentation method, a dual morphological grayscale reconstruction method is used to remove noise and highlight nuclear shapes. The initial nucleus boundaries are obtained based on the marker-based watershed algorithm. Then, a snake model is used to segment the shapes of nucleus regions precisely. We previously proposed a similar approach for segmentation of cervical cells in [19]. However, obtaining a corresponding marker for each nucleus region is a problem due to the variable nature of overlapping cells.

2.2 Classification of Cervical Cells

In the literature, there are many studies related to classification of cell images. Huang and Lai [16] propose an SVM-based decision graph classifier for classification of hepatocellular carcinoma biopsy images. Walker et al. [38] use a quadratic Bayesian classifier to classify nuclei of cervical cells based on their textural features. Theera-Umpon [36] use neural networks for classifying white blood cell images. Bazoon et al. [7] utilizes a hierarchical system of artificial neural networks using back-propagation for classification of cervical cells.

In this work, we concentrate on the ranking of cervical cells rather than their classification for a number of reasons. First, classification requires a large training set containing the complete repertoire of expected cell patterns for each class. Collecting such a training set is a very challenging task and entails a long period of time because cells on two slides may be quite different from each other due to artifacts, overlapping cells, and inconsistent staining. Moreover, the complexities of cellular analysis and the need for high sensitivity and specificity make human intervention inevitable. Two semi-automated slide scanning devices approved by the FDA in the USA retrieve fields of diagnostic interest for examination of cytotechnicians rather than classifying slides [12]. Lastly, we aim to approach the problem in a different way by proposing an unsupervised screening system.

There is no such a system automating Pap smear screening by computers without human intervention [12]. However, two semi-automated commercial devices, namely the FocalPoint GS Imaging system and the ThinPrep Imaging system, are used for interpreting Pap slides. The FocalPoint system contains three cameras with 4x and 20x magnifications. The scanning cameras measure over 300 different features like size, texture, density, cytoplasmic features, shape features, nucleus/cytoplasm ratio and so on. The system retrieves 10 fields of diagnostic interest in the ranked order. The ThinPrep system differs from the FocalPoint system in the sense that it retrieves 22 fields of interest without ordering. Hence, all of the retrieved slides need human intervention. Note that the imaging capability of these devices provides numerous advantages over our imaging system consisting of a microscope connected to a digital camera.

In this thesis, we propose to order cells in a Pap smear slide according to their abnormality degree. Ranking is an important task in information retrieval area where the items are ordered based on their similarity to a reference item called query. We cannot employ this approach easily because our problem involves no query cell. Hence, we first perform unsupervised hierarchical clustering on the cell features and obtain an initial ordering of the cells as the leaf ordering of the hierarchical tree. We further improve this ordering by applying the fast optimal leaf ordering algorithm [6].

Hierarchical clustering has been extensively used in biological literature to explore related genes that share a common function [6, 15, 34, 2, 8]. Biological analysis is often done in the context of the linear leaf ordering of the hierarchical tree. Thus, finding a suitable ordering of the leaves consistent with the tree structure is studied in the literature. Eisen et al. [15] order leaves of a hierarchical tree based on their average expression level. Tamayo et al. [35] propose to order leaves using the results of a one dimensional self organizing map. Alon et al. [3] order leaves and internal nodes based on their similarity to parent's siblings. We use the fast optimal leaf ordering algorithm because all of the above methodologies are based on heuristics. The optimal leaf ordering algorithm searches all possible orderings of the leaves in order to find an optimal one that maximizes a criterion function such as the sum of similarities between adjacent leaves.

Chapter 3

Segmentation of Cervical Cells

In this work, we propose a computer-assisted screening system which orders cells in an image of a Pap smear slide according to their dysplasia degree such that the cells ranked as more normal than a selected cell which is manually identified as normal can be skipped by cyto-technicians. Evaluation of a cell is guided by the measurements of geometric properties of nucleus and cytoplasm such as area, radius, perimeter, convexity, etc. Thus, the first and most crucial step of the proposed system is the accurate segmentation of cells along with their nucleus and cytoplasm.

There are common problems encountered in Pap smear images similar to the ones existing in other immunohistochemically stained cytological images. Firstly, all parts of cells may not be equally stained by traditional staining techniques and this causes inhomogeneity in a single slide and inconsistency between different slides. Moreover, cells are usually grouped and they may overlap or occlude each other. It is a very difficult task to differentiate boundaries of overlapping or occluding cells even manually. Figure 3.1 shows an example Pap smear image illustrating these characteristic problems.

In this thesis, a three-phase approach to segmentation is used where thresholding method is used as the first phase to extract background regions for obtaining remaining cell regions. The second phase consists of segmenting the cell regions

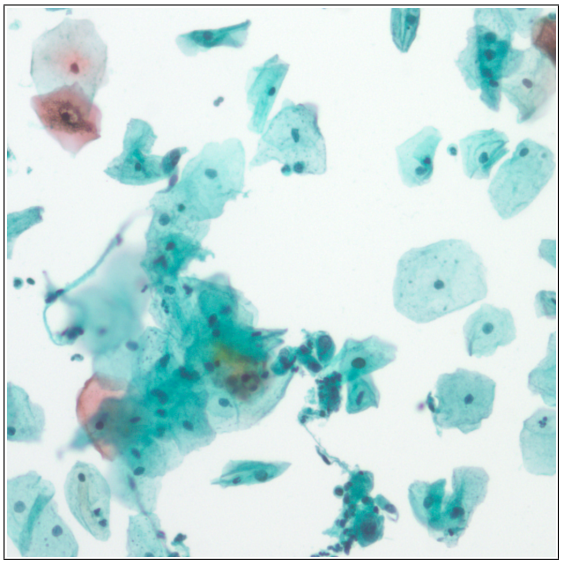


Figure 3.1: An example Pap smear image with its characteristic problems such as inhomogeneous staining, overlapping cells.

by a non-parametric hierarchical segmentation algorithm that uses the spectral and shape information as well as the gradient information. The last phase aims to partition the cell region into true structures of each nucleus and the whole cytoplasm area by classifying the final segments as nucleus or cytoplasm region. The details of each step are explained in the following sections.

3.1 Background Extraction

Cervical cells on a Pap smear slide are colored with the tones of blue and red colors as a result of the staining process. The remaining arbitrary empty background regions that do not include any cytological structures remain colorless, and produce white pixels. Cell and background regions have distinctive colors in terms of brightness such that they can be distinguished according to their gray level values of luminance.

Pap smear images are obtained from a microscope with the help of a camera and they are initially in the RGB color space. We convert the images from the

RGB color space to the Lab color space developed by the International Commission on Illumination (CIE). The Lab is a perceptually uniform color space meaning that a change of the same amount in a color value should produce the same amount of perceptual difference of visual importance [28]. L channel corresponds to illumination and, a and b channels correspond to the color opponent dimensions. It is derived from the CIE XYZ color space, which is based on direct measurements of human visual perception.

Our goal using the Lab color space is to separate color and illumination information and analyze the histogram of L channel which represents brightness measure. Figure 3.2 shows an example Pap smear image in RGB color space, L channel of the transformed image in CIE Lab color space and the corresponding histogram of the L channel. At the end of this section, we further explain the reason that L channel of the Lab color space is analyzed for separating the background and cell regions.

Pap smear images usually have non-homogeneous illumination due to uneven lightening of the slides during image acquisition as illustrated in Figure 3.3. Since the cells are darker than the background, we use the black top-hat transform for mitigating inhomogeneous illumination. The black top-hat or top-hat by closing BTH of an image I is defined as the difference between the closing $(I \bullet SE)$ of the original image by a structuring element SE and the original image [18]:

$$BTH = (I \bullet SE) - I. \tag{3.1}$$

Figure 3.3 (b) shows that closing with a disk structuring element of radius 210 removes the cells but preserves the illumination function. As illustrated in Figure 3.3 (c), the subtraction of the image from the closing of it provides us with an evenly illuminated image which we call the illumination-corrected L channel in the rest of the thesis. Note that the cells become lighter than the background in the illumination-corrected L channel because of the black top-hat transform.

It is very hard to infer a threshold between the gray level values of the background and cell regions by considering the histogram of the L channel in Figure 3.2 (c). However, it is appropriate to assume that the respective populations of background and cell regions are distributed normally with distinct means and

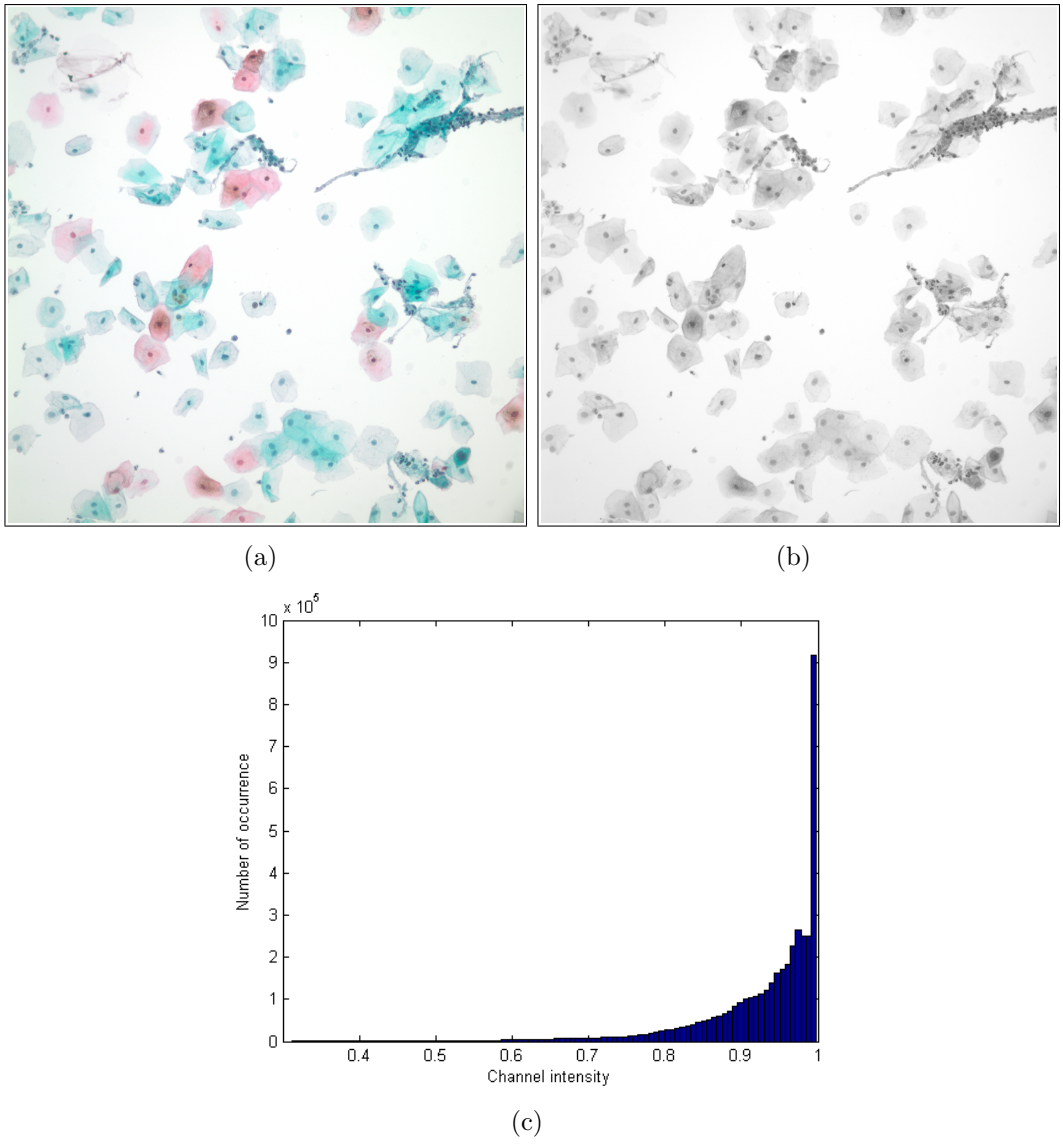


Figure 3.2: A Pap smear image (a) in RGB color space (b) L channel of the transformed image in CIE Lab color space (c) the histogram of the L channel.

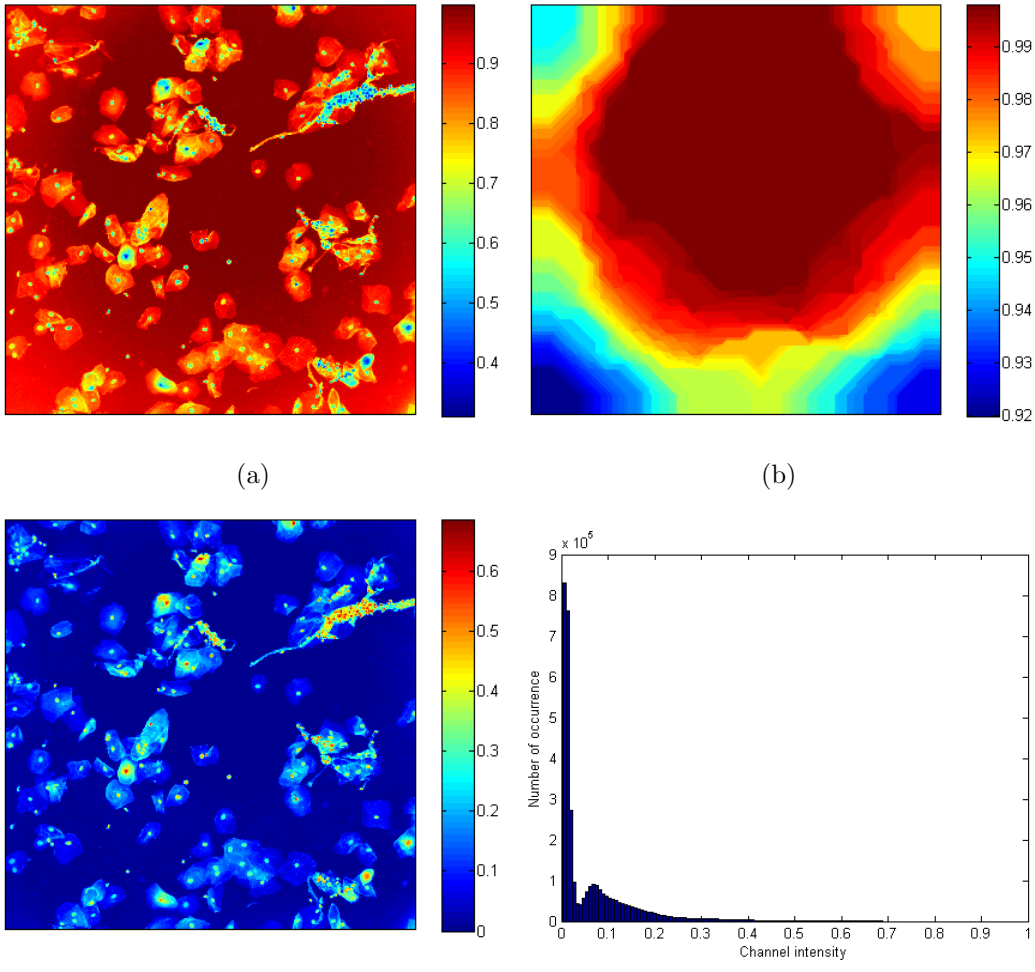


Figure 3.3: Use of black top-hat transform for mitigating inhomogeneous illumination (a) L channel of the image (b) closing with a large structuring element (c) the illumination-corrected L channel by the black top-hat transform (d) the histogram of the illumination-corrected L channel.

standard deviations based on the corresponding histograms of Pap smear images constructed after illumination correction as in Figure 3.3 (d). In this work, a suitable threshold between the gray level values of background and cell regions is determined by the minimum error thresholding method of Kittler and Illingworth [22] as explained below.

Let us consider an image whose pixels have gray level intensity values x from the interval [0, n]. The histogram of gray level image, $h(x)$, gives the frequency of occurrence of each gray level x in the image. Thus, $h(x)$ is the estimate of the probability density function of the mixture population composed of gray levels of background and objects. Suppose that we threshold the gray level data at some arbitrary level T and model each of the two resulting pixel populations by normal density model $p(x|\omega_i, T)$ with mean μ_i, standard deviation σ_i and a priori probability $p(\omega_i|T)$ given as

$$p(\omega_i|T) = \sum_{x=a}^{b} h(x), \tag{3.2}$$

$$p(x|\omega_i, T) = \frac{1}{\sqrt{2\pi}\,\sigma_i}\, exp\left(-\frac{(x-\mu_i)^2}{2{\sigma_i}^2}\right), \tag{3.3}$$

where ω_1 and ω_2 corresponds to object and background, respectively, and

$$\mu_i = \left(\sum_{x=a}^{b} x\, h(x)\right) / p(\omega_i|T), \tag{3.4}$$

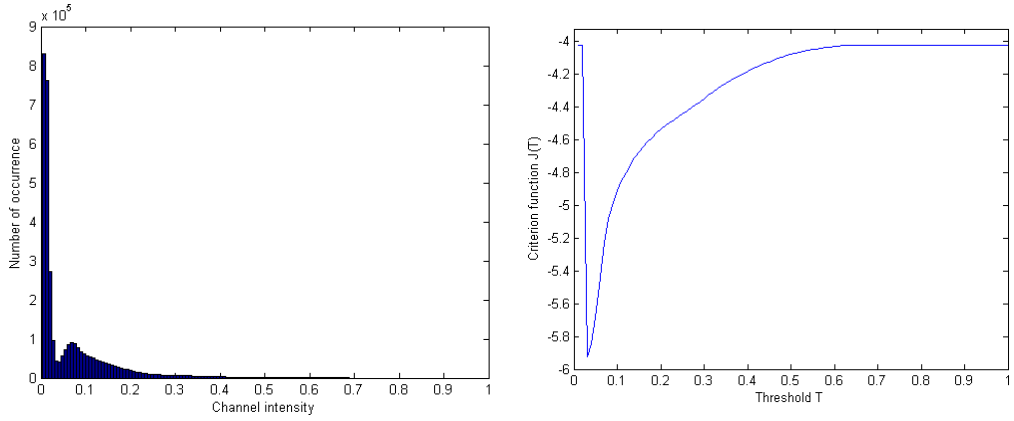
$${\sigma_i}^2 = \left(\sum_{x=a}^{b} (x-\mu_i)^2 h(x)\right) / p(\omega_i|T), \tag{3.5}$$

$$a = \begin{cases} 0, & \text{i} = 1, \\ T+1, & \text{i} = 2, \end{cases} \tag{3.6}$$

$$b = \begin{cases} T, & \text{i} = 1, \\ n, & \text{i} = 2. \end{cases} \tag{3.7}$$

The posterior probability $p(\omega_i|x, T)$ of gray level x being classified correctly is given by

$$p(\omega_i|x, T) = \frac{p(x|\omega_i, T)\, p(\omega_i|T)}{h(x)}, \qquad i = \begin{cases} 1, & \text{x} \leq \text{T}, \\ 2, & \text{x} > \text{T}. \end{cases} \tag{3.8}$$



(a) (b)

Figure 3.4: (a) The histogram taken from Figure 3.3 (b) its corresponding criterion function.

As $h(x)$ is independent of both i and T, the denominator in (3.8) is ignored in the analysis. Taking the logarithm of the numerator in (3.8) and multiplying the result by -2, we obtain

$$\epsilon(x,T) = \frac{(x-\mu_i)^2}{{\sigma_i}^2} + 2\,\log(\sigma_i) - 2\,\log(p(\omega_i|T)), \qquad i = \begin{cases} 1, & \text{x} \leq \text{T}, \\ 2, & \text{x} > \text{T}, \end{cases} \tag{3.9}$$

which can be considered as an alternative index of correct classification performance. The average performance of the thresholding can be measured by the criterion function

$$J(T) = \sum_x h(x)\,\epsilon(x,T). \tag{3.10}$$

The distribution models of the background and object populations change according to the selected threshold T. The criterion indicates indirectly the amount of overlapping between the Gaussian models of these populations. When the models and the data fit better, the overlap between density functions decreases. Thus, a smaller value of criterion function means smaller classification error.

The criterion function can be expressed as

$$J(T) = \sum_{x=0}^{T} h(x) \times \left(\frac{(x-\mu_1)^2}{{\sigma_1}^2} + 2\,\log(\sigma_1) - 2\,\log(p(\omega_1|T)) \right) + \sum_{x=T+1}^{n} h(x) \times \left(\frac{(x-\mu_2)^2}{{\sigma_2}^2} + 2\,\log(\sigma_2) - 2\,\log(p(\omega_2|T)) \right). \quad (3.11)$$

Substituting (3.2) through (3.5) into (3.11), we find

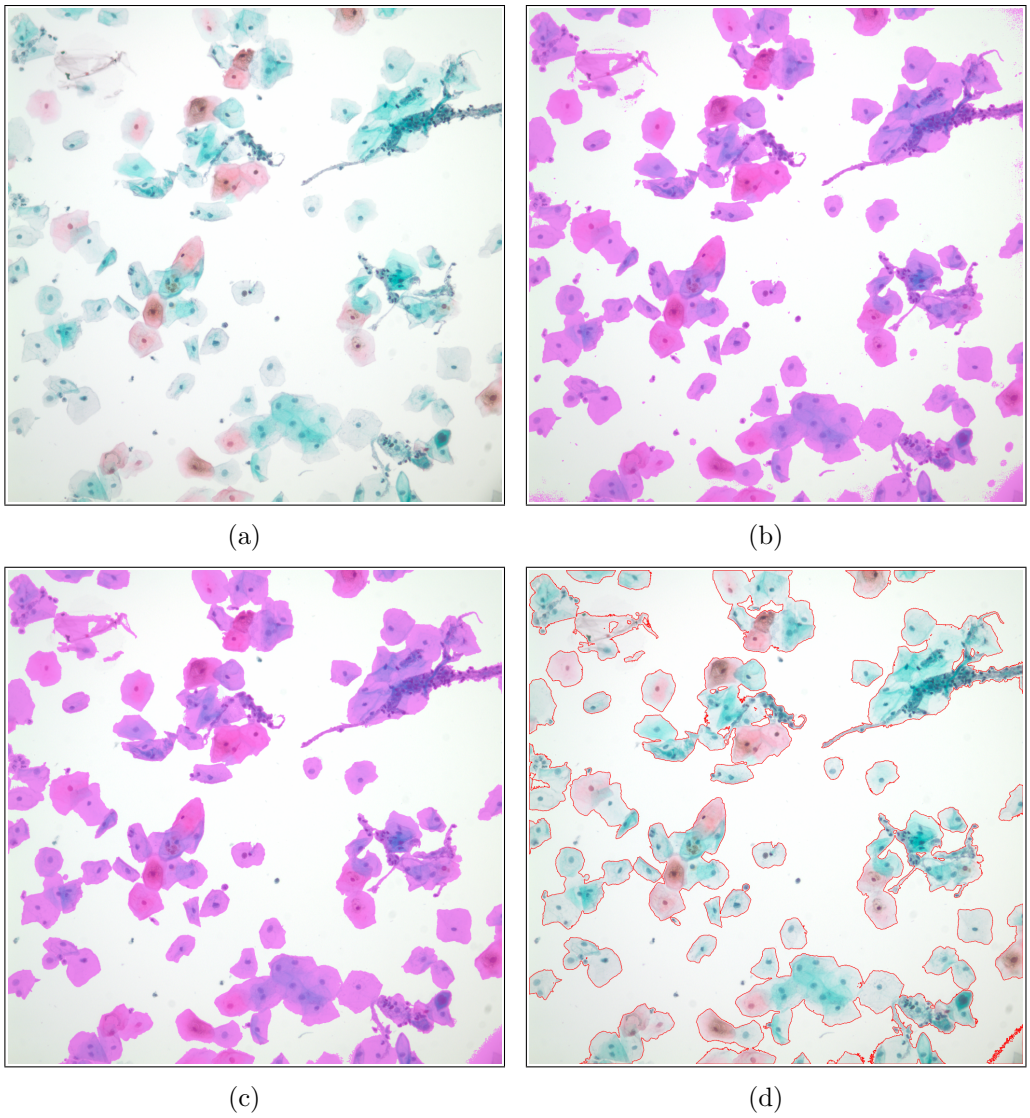
$$J(T) = 1 + 2\,[\,p(\omega_1|T)\,\log(\sigma_1) + p(\omega_2|T)\,\log(\sigma_2)\,] - 2\,[\,p(\omega_1|T)\,\log(p(\omega_1|T)) + p(\omega_2|T)\,\log(p(\omega_2|T))\,]\,. \quad (3.12)$$

The optimal threshold can be found by minimizing the criterion function expressed in (3.12).

The histogram taken from Figure 3.3 and its corresponding criterion function for different values of threshold T are shown in Figure 3.4. The unique global minimum of the criterion function implies histogram bimodality and the optimal threshold is found as 0.03. The pink areas shown in Figure 3.5 (b) are the candidate cell regions obtained by thresholding the image in Figure 3.3 (c) according to the optimal threshold. The last step includes elimination of the regions whose area is smaller than the smallest possible cell size which is empirically determined as 1500 pixels for Pap smear images taken at magnification 20x. The cell regions are colored with pink in Figure 3.5 (c) and their boundaries are red in Figure 3.5 (d).

In order to segment the Pap smear images into the background and cell regions, we first transform the images from the RGB color space to the Lab color space. Then, the illumination correction is applied on the L channel for filtering non-homogeneous illumination. Lastly, we determine the threshold between background and cell regions by using minimum error thresholding.

Our experiments showed that the background and cell regions can be distinguished according to their brightness measure which is represented by the illumination-corrected L channel. We also analyze the histograms and segmentation results obtained from other color spaces. Figure 3.6, Figure 3.7 and Figure



(a) (b)

(c) (d)

Figure 3.5: (a) The Pap smear image taken from Figure 3.2 (b) regions found by thresholding at $T = 0.03$ (c) cell regions after eliminating small areas (d) boundaries of cell regions.

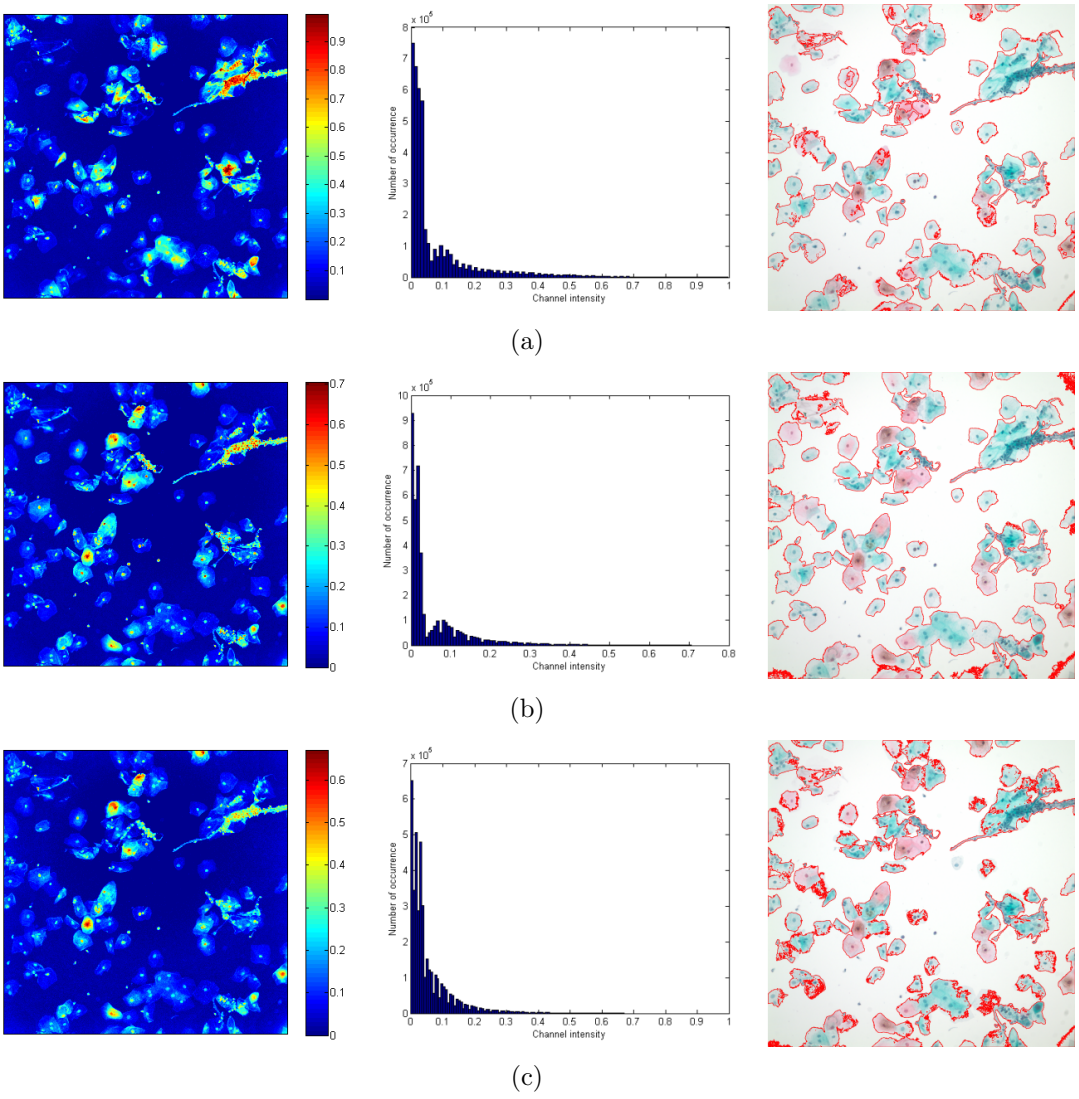


Figure 3.6: Illumination-corrected channels of RGB color space, their histograms and segmentation results (a) R channel (b) G channel (c) B channel.

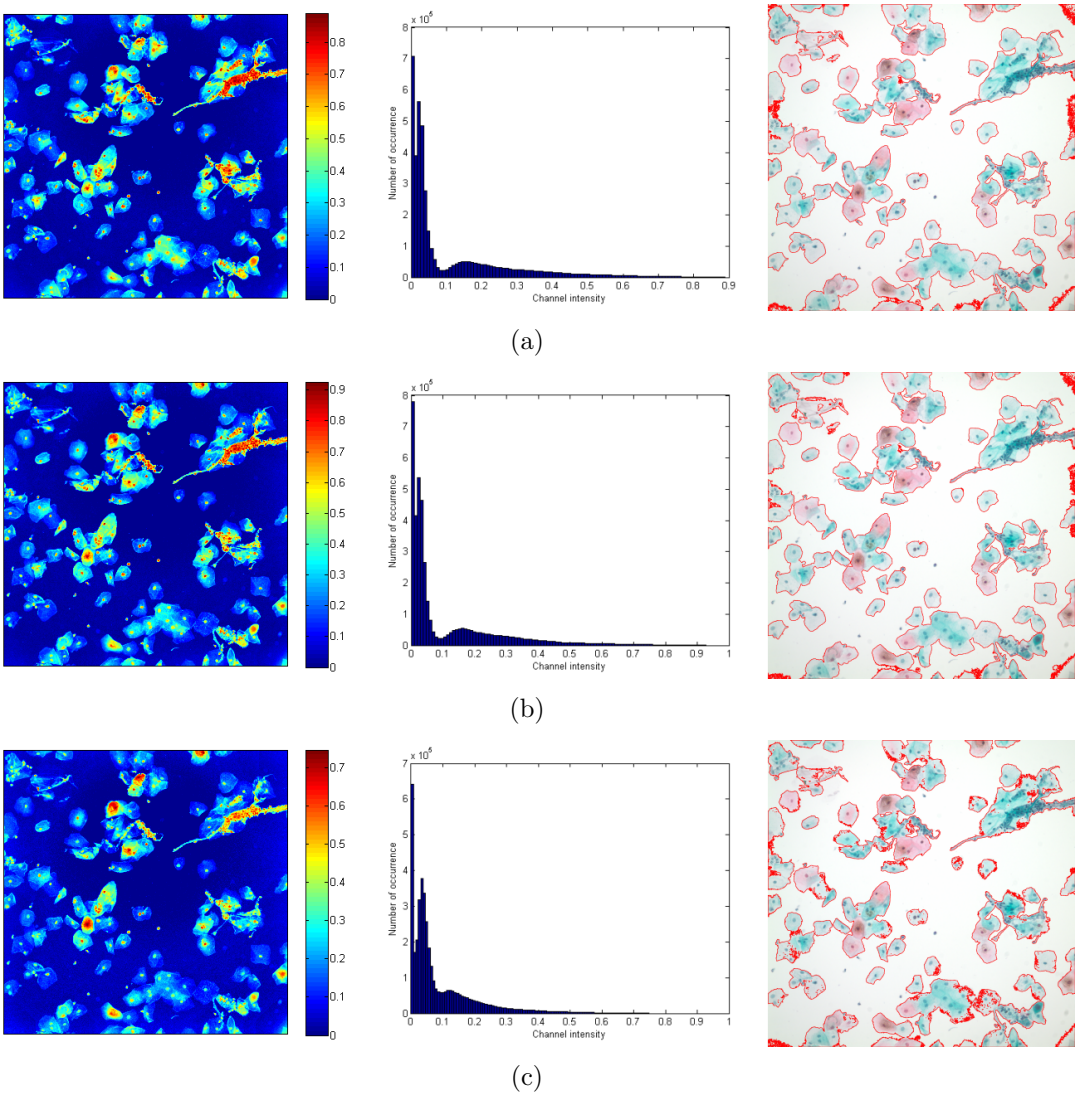


Figure 3.7: Illumination-corrected channels of CIE XYZ color space, their histograms and segmentation results (a) X channel (b) Y channel (c) Z channel.

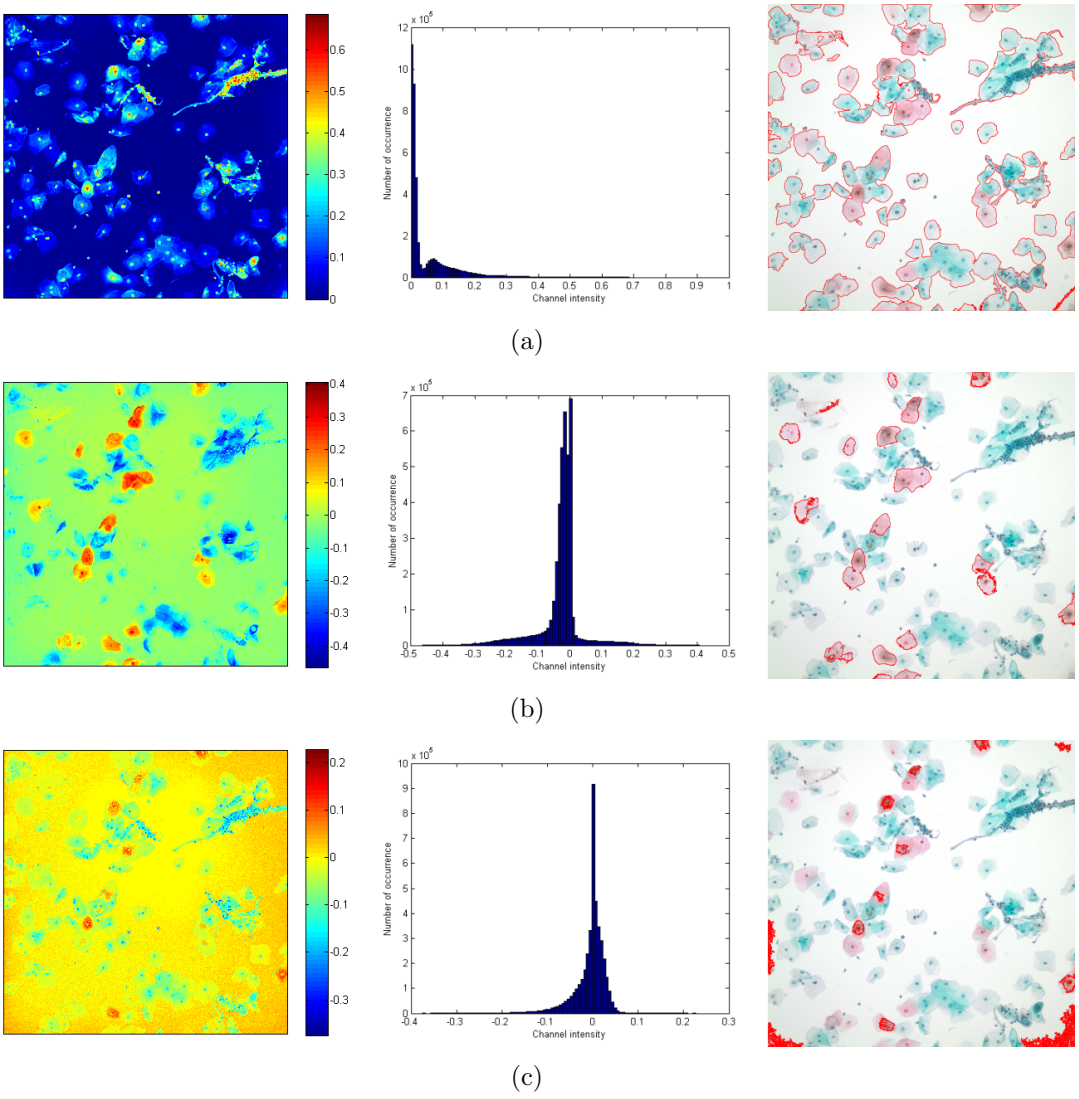


Figure 3.8: Channels of CIE Lab color space, their histograms and segmentation results (a) illumination-corrected L channel (b) a channel (c) b channel.

3.8 show each channel of RGB, CIE XYZ, and CIE Lab color spaces after illumination correction, their corresponding histograms, and segmentation results, respectively, where the boundaries of the cell regions are colored with red. We get better segmentation result when the structure of the histogram is more bimodal because minimum error thresholding method is based on the assumption that background and object populations are distributed normally with distinct means and standard deviations. For example, the corresponding segmentation results of the channels a and b are not correct due to the missing bimodal structure in the histograms of these channels. Moreover, when the estimated Gaussian models of the populations and the data fit better, the overlap between density functions decreases and the segmentation results improve. These two conditions are satisfied better for the illumination-corrected L channel of Lab color space from which the best segmentation is obtained.

3.2 Nucleus and Cytoplasm Segmentation

In this section, our goal is to segment the cell regions that are obtained in the previous step into the areas of nucleus and cytoplasm. There are many studies [40, 41] on the segmentation of cell images in the literature, and some of them [24] include specific methods for the images of immunohistochemically stained cytology specimens. However, the cell segmentation remains a problem due to the complex and variable nature of cell structures with the inconsistency between different images resulting from the staining process. Variability of image intensity exists even in a single cell, and this makes it difficult to use thresholding methods for segmentation. Edge based approaches assume that discontinuities in the image intensity imply boundaries between different objects. Clustering and histogram based segmentation algorithms rely on the notion that images have a reasonable number of objects with homogeneous features. The performance of these methods is substantially affected by the noise and artifacts frequently encountered in the cell images. On the other hand, region based approaches involving growing, splitting, and merging of regions are robust to noise [32].

In a related work, Akçay and Aksoy [1] develop a segmentation method that uses the neighborhood and spectral information as well as the morphological information. They first extract candidate regions by applying morphological opening and closing by reconstruction operations. Then, a hierarchical tree is constructed from the extracted regions, and the most meaningful regions in that tree are selected by optimizing a measure consisting of two factors: spectral homogeneity, and neighborhood connectivity. Spectral homogeneity is calculated in terms of variances of multi-spectral features, and neighborhood connectivity is calculated in terms of sizes of connected components.

In this work, we propose to segment the cell regions into the nucleus and cytoplasm areas by following up the general framework of the segmentation method developed by Akçay and Aksoy [1]. The main difference and advantage of our approach stems from being a non-parametric hierarchical segmentation algorithm that uses the spectral and shape information as well as the gradient information. In their work [1], Akçay and Aksoy apply their segmentation method on remotely sensed images in which they aim to find meaningful objects like buildings, roads, vegetation, etc. The properties of images and objects we deal with are different from the properties of remotely sensed images and objects in them. Hence, instead of using morphological opening and closing by reconstruction operations, we extract the candidate regions by applying watershed segmentation to h-minima transforms of the image gradient for increasing values of h. Then, we similarly construct a hierarchical tree from the extracted regions and select the most meaningful regions in that tree by optimizing a measure. However, the measure we use is different from the one used in [1] such that our measure consists of two factors as the spectral homogeneity, which is calculated in a different way, and the circularity. Finally, we classify the selected regions as nucleus or cytoplasm regions according to their size, mean intensity, circularity, and homogeneity features.

3.2.1 Hierarchical Region Extraction

In this section, we compare three different methods for extracting candidate regions to construct the hierarchical tree. The first method is the one used in [1]

which is based on morphological opening and closing operations. The remaining two methods can be considered as the multi-scale watershed segmentation based on dynamic concept associated to regional minima. The difference between these two methods is that the second method uses minima imposition technique whereas the third method is based on the h-minima transform. We describe each of three methods along with their advantages and drawbacks below.

In the related work of Akçay and Aksoy [1], the candidate segments are found by using opening and closing by reconstruction operations. Opening by reconstruction (respectively, closing by reconstruction) operation preserves the shape of the structures that are not removed by erosion (respectively, dilation). They first calculate the morphological profiles by applying opening and closing by reconstruction operations using increasing structuring element (SE) sizes. Then, the derivative of the morphological profile (DMP) is used to find the candidate segments which are composed of a neighboring group of pixels with a similar change for any particular SE size. The DMP [29] is defined as a vector where measure of the slope of the opening-closing profile is stored for every step of an increasing SE series. They assume that pixels with a positive DMP value for a particular SE size have a similar change with respect to their neighborhoods at that scale. They obtain final candidate segments by applying connected components analysis to the DMP at each scale.

Figures 3.9 and 3.10 show the red colored boundaries of the candidate regions obtained for different scales by the morphological profiles of closing by reconstruction and opening by reconstruction operations, respectively, using disk shaped structuring elements (SE size corresponds to disk radius). Note that many segments at the smaller scales are appearing due to the heterogeneous and textured structure of the nucleus and cytoplasm areas. At the later scales, segments corresponding to true structures like the nucleus regions start to merge with their surroundings and other components after reaching the SE size corresponding to the radius of a disk structuring element in which they appear. Moreover, the cytoplasm region does not have a regular geometric structure and its size depends on the number and type of the occluding cells in it. Thus, it is difficult to obtain a candidate segment related to the whole cytoplasm. The nucleus regions are in

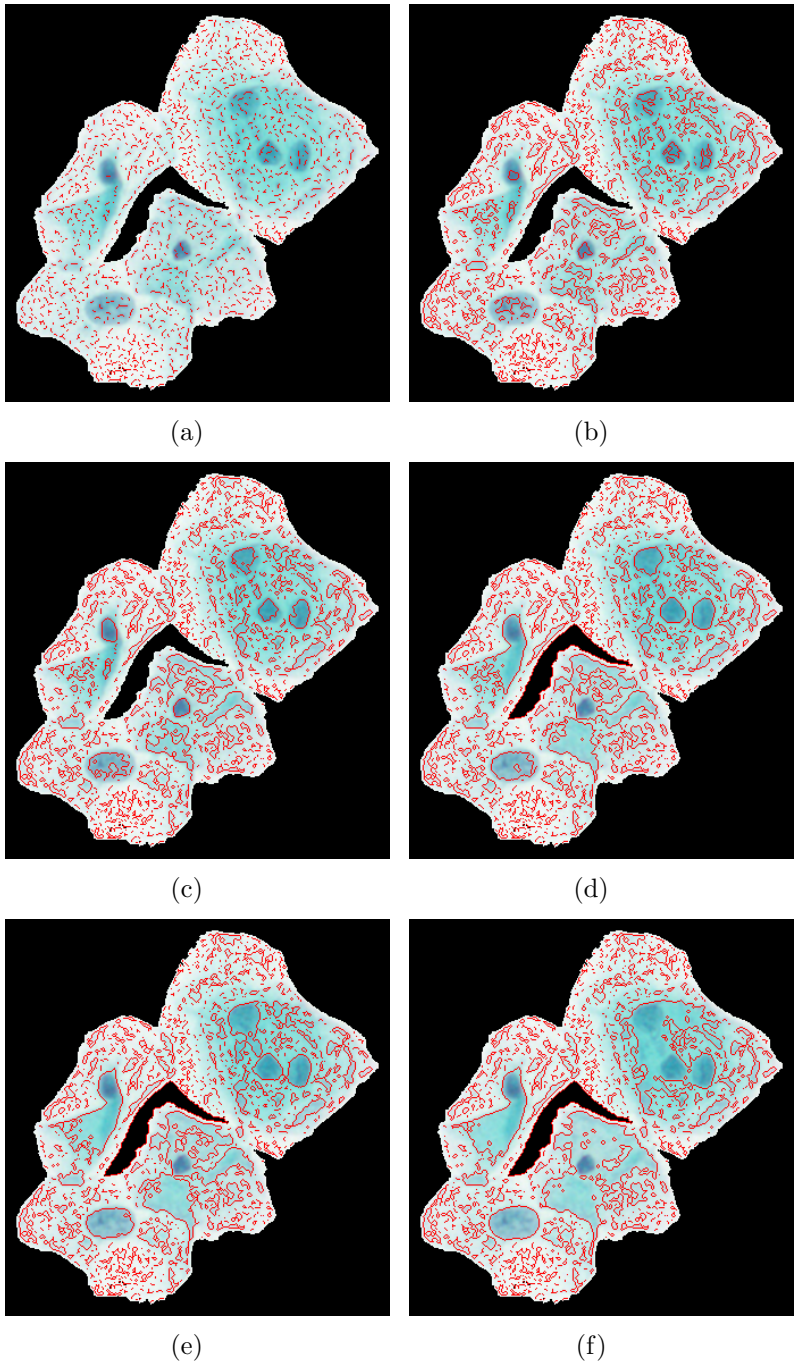


Figure 3.9: Candidate segments obtained by morphological profiles of closing by reconstruction (a) at SE size 1 (b) at SE size 4 (c) at SE size 7 (d) at SE size 10 (e) at SE size 13 (f) at SE size 15.

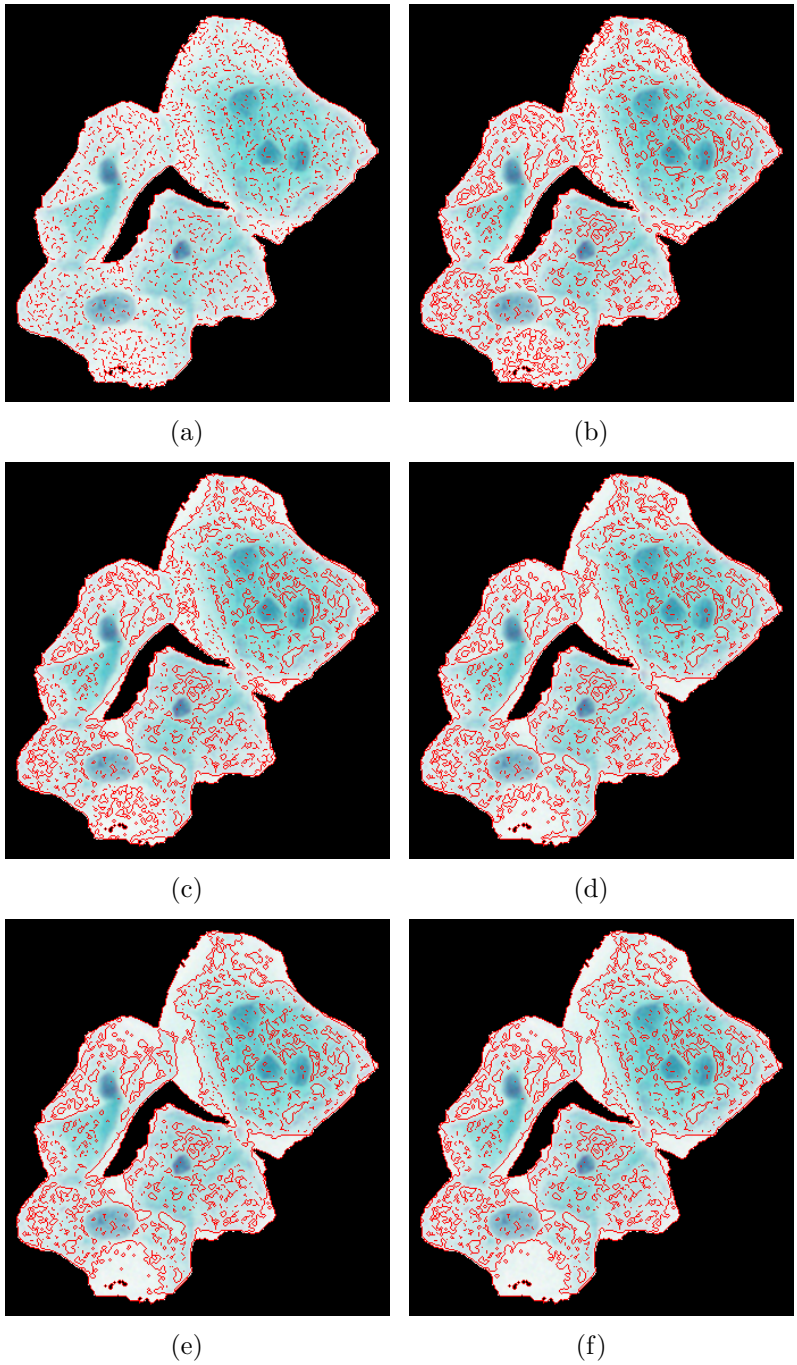


Figure 3.10: Candidate segments obtained by morphological profiles of opening by reconstruction (a) at SE size 1 (b) at SE size 4 (c) at SE size 7 (d) at SE size 10 (e) at SE size 13 (f) at SE size 15.

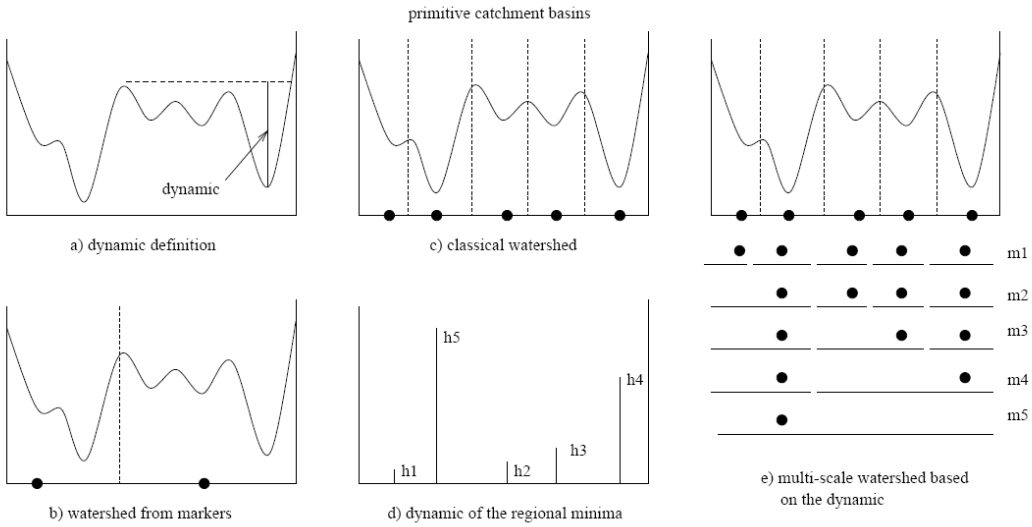


Figure 3.11: Hierarchical watershed transform based on dynamics. (Image taken from [13].)

a regular elliptic shape but their size differs according to their cell type. When the number of scales in the hierarchy is increased, candidate segments related to nuclei merge with their surroundings originated from the heterogeneous and textured cytoplasm area. Selecting a suitable size for the largest SE is important to ensure that candidate segments for all nuclei are generated but they are not allowed to merge with their surrounding noisy segments much.

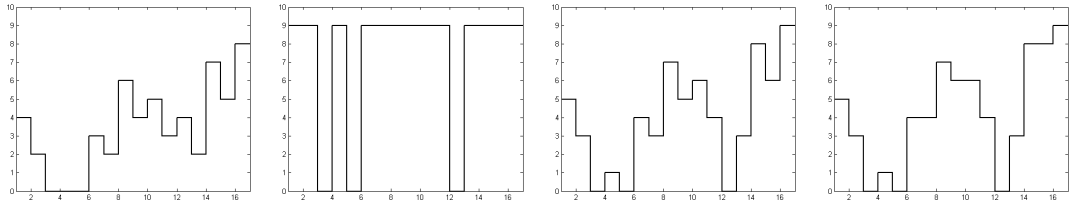
The second method for generating a hierarchical partitioning of the input image is the multi-scale watershed segmentation based on the concept of dynamic related to regional minima. A regional minimum is a connected component of pixels with a single intensity value t whose external boundary pixels have a value strictly greater than the value t. When we consider the input image as a topographic surface, the dynamic or depth of a regional minimum becomes the minimum height that a point in the minimum has to climb to reach a lower regional minimum. Figure 3.11 (a) illustrates the dynamic of a regional minimum in a one-dimensional signal and Figure 3.11 (d) shows the dynamic values of all regional minima in the signal.

We first review the watershed segmentation and its marker-controlled counterpart for completeness. Watershed concept comes from the field of topography. When we consider the gradient of an image as a topographic surface, the high gradient image edges correspond to the watershed lines and the low gradient regions correspond to the catchment basins. The pixels that drain to the same regional minimum belong to the same catchment basin. Each pair of catchment basins are separated by a watershed line and the union of all watershed lines defines the watershed segmentation. The watershed segmentation of the example signal is given in Figure 3.11 (c).

Watershed segmentation can be simulated by an immersion process. If we immerse the topographic surface associated with the image gradient in water, the water rises through the holes at regional minima with a uniform rate. When the water coming from two different minima is about to merge, a dam is built at each point of contact. Following the immersion process, the union of all those dams constitutes the watershed lines. Efficient algorithms implementing the watershed segmentation are proposed in the literature [37, 25].

Marker controlled watershed segmentation can be defined as the watershed of an input image transformed to have regional minima only at the marker locations. We rearrange the image minima by using minima imposition technique based on a set of markers marking relevant objects. Figure 3.12 illustrates the steps of the minima imposition on a one-dimensional signal. The marker image f_m consists of pixels whose value is 0 at the marker locations and t_{max} at the rest of the image. First, we create minima only at the locations of markers by taking the point-wise minimum between $f+1$ and f_m. Note that the resulting image is lower or equal to the marker image. The second step of the minima imposition is the morphological reconstruction by erosion of the resulting image from the marker image f_m. Figure 3.11 (b) illustrates the marker locations and the watershed segmentation of the signal from these markers.

The multi-scale watershed segmentation generates a set of nested partitions where each partition is obtained by applying marker controlled watershed to the input image using a decreasing set of markers. The partition at scale s is



(a) (b) (c) (d)

Figure 3.12: (a) One-dimensional signal f (b) marker f_m (c) point-wise minimum between $f + 1$ and f_m (d) reconstruction by erosion of (c) from f_m.

the marker controlled watershed segmentation of the image using the markers whose locations are determined as the regional minima having a dynamic greater than s. The partition at scale 0 is the classical watershed made of primitive catchment basins. As the scale increases, fewer markers are involved and the coarsest partition is the entire image obtained from the single marker with the largest dynamic. Figure 3.11 (e) shows the hierarchical watershed partitioning of a signal. There are five regional minima with dynamics from $h1$ to $h5$ and five different partitions $m1$ to $m5$ where the regional minimum with the next smallest dynamic is suppressed in each partition. $m1$ is the classical watershed with five catchment basins. At scale 2, left primitive catchment basin having the smallest dynamic is merged to its neighbor catchment basin and $m2$ has four catchment basins. $m4$ has the two most relevant catchment basins.

Figure 3.13 shows the pairs of consecutive partitions obtained for the example cell image at different scales such that the second partition of each pair is calculated by suppressing the regional minima of the dynamic value less than or equal to the smallest dynamic value of the first partition. For example, we obtain the partition at scale 14 by applying watershed segmentation to the image gradient after suppressing its regional minima with a dynamic value less than or equal to 14 which is the smallest dynamic value of the partition at scale 13.

Note that true structures of some nuclei are obtained in the partitions of later scales like the segments associated to three nuclei at scale 13 and this observation confirms that the nucleus regions are associated with higher dynamic values.

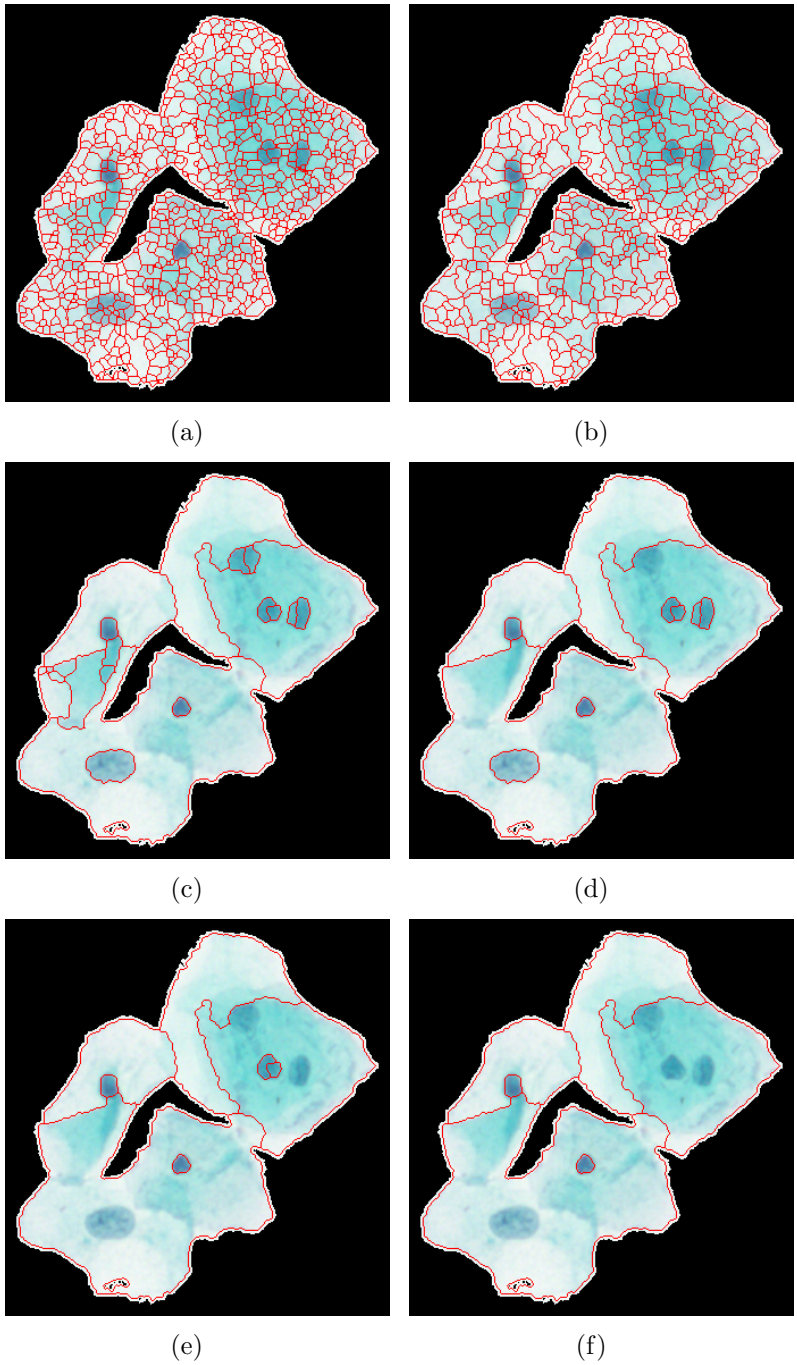


Figure 3.13: Candidate segments obtained by multi-scale watershed segmentation based on dynamic. (a) at scale 0 (b) at scale 1 (c) at scale 13 (d) at scale 14 (e) at scale 26 (f) at scale 27.

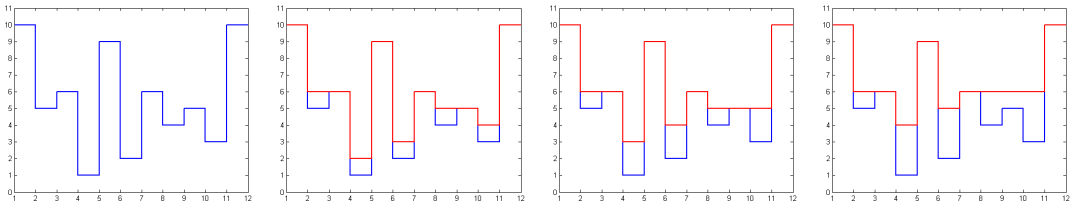
However, the segments constituting some nucleus regions such as the upper nucleus and the one below it merge with the surrounding cytoplasm region at the same time without forming the true structure of the nucleus. For example, the two segments constituting the upper nucleus at scale 13 merge with the cytoplasm region at the next scale 14 without merging with each other to form a segment of the whole nucleus. Similarly, the segments constituting the nucleus below the upper one at scale 26 merge with the cytoplasm region at the next scale 27. This implies that the segments constituting each of these two nuclei have the minima whose highest dynamic value is the same.

Another method for extracting candidate regions to construct the hierarchical tree is the multi-scale watershed segmentation based on h-minima transform. For completeness, we first give an explanation of the h-minima transform below.

H-minima transform suppresses all minima whose dynamic or depth is lower than or equal to a given threshold h [33]. This is achieved by performing geodesic reconstruction by erosion of the input image f from $f + h$. Figure 3.14 shows a one-dimensional signal (blue) and its h-minima transformations (red) for different values of h. We can observe that the minima of depth lower than or equal to h are filtered whereas the other minima of depth higher than h either remain same or are extended as shown in Figure 3.14 (c).

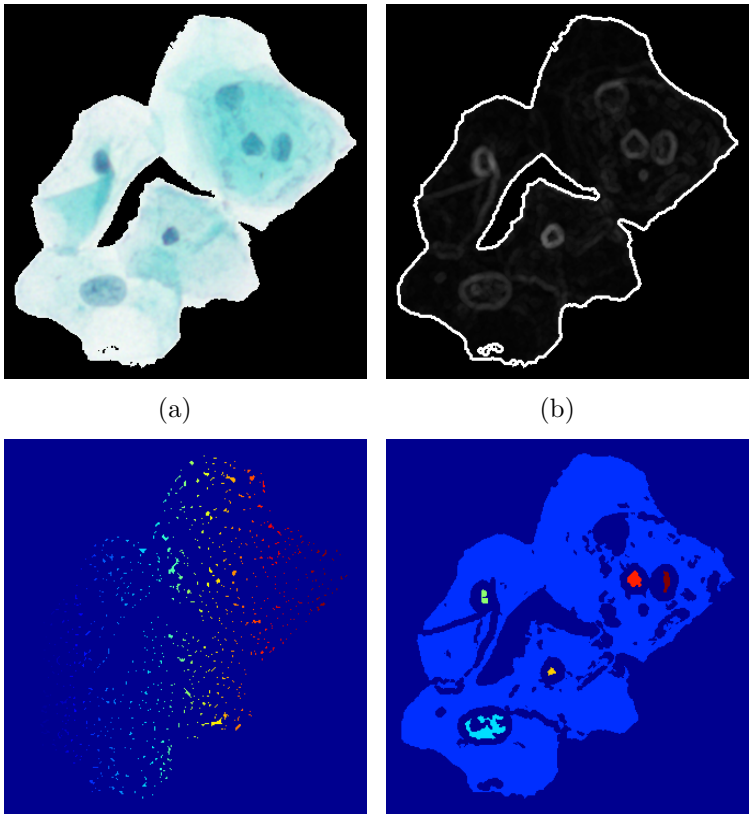
We illustrate how the h-minima transform changes the minima located at the gradient of the example cell image in Figure 3.15. The regional minima of the raw image gradient are shown in Figure 3.15 (c). These minima are mainly marking the texture occurring in the nucleus and cytoplasm regions because no pre-filtering is applied. Figure 3.15 (d) presents the minima of the h-minima transform of the image gradient for $h = 17$. These minima better mark relevant dark nucleus structures of the input image.

We calculate a set of nested partitions of cell images by applying the watershed segmentation to the h-minima transform of the image gradient for increasing values of h. The watershed partition at scale s is the watershed of the image gradient whose regional minima of depth less than or equal to s are suppressed by the h-minima transform. Similar to the second method, the partition of scale



(a) (b) (c) (d)

Figure 3.14: (a) One dimensional signal (blue) (b) h-minima transformations (red) for $h = 1$ (c) $h = 2$ (d) $h = 3$.



(a) (b)

(c) (d)

Figure 3.15: (a) Cell image (b) its gradient (c) the minima at the raw gradient (d) the minima at the h-minima of the gradient for $h = 17$.

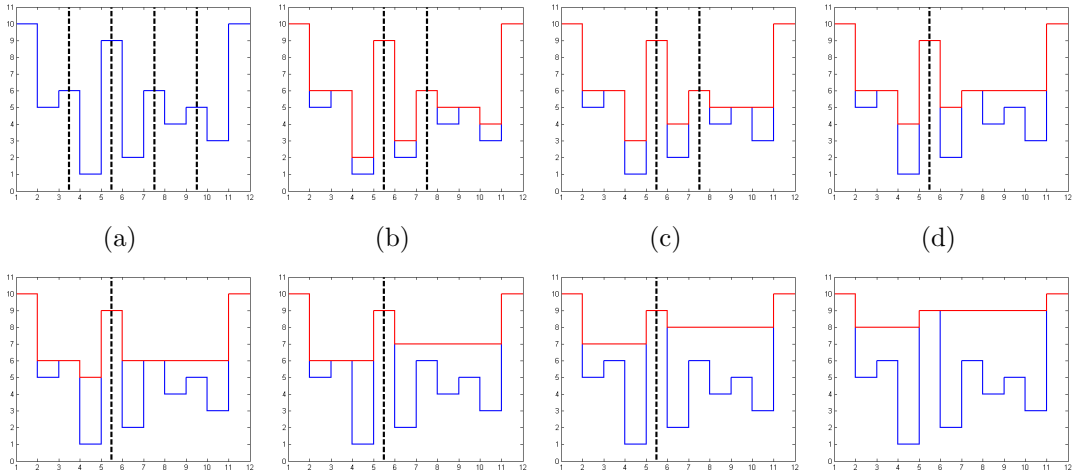


Figure 3.16: One dimensional signal (blue) and watersheds (black) of h-minima transformations (red) (a) at scale 0 (b) at scale 1 (c) at scale 2 (d) at scale 3 (e) at scale 4 (f) at scale 5 (g) at scale 6 (h) at scale 7.

0 is the classical watershed made of primitive catchment basins. As the scale increases, more regional minima are filtered and the coarsest partition is the entire image obtained from a single regional minimum of the largest depth.

Figure 3.16 shows the hierarchical watershed partitioning of the one-dimensional signal at different scales. There are 4 different partitions of the signal observed at 8 scales. The first partition is the classical watershed with five catchment basins. At scale 1, the two primitive catchment basins having the smallest dynamic value are merged to their neighbor catchment basins and the second partition has three catchment basins. The partition of scale 3 has two most relevant catchment basins.

An example of the multi-scale watershed partitioning based on the h-minima transform applied to the cell image is shown in Figure 3.17. Parts (a) through (f) show the six levels in the hierarchy calculated from the gradient image which is transformed by suppressing the minima whose dynamics are lower than or equal to 2, 3, 6, 7, 12, and 13, respectively. The true structures of nucleus segments are better obtained in the third method compared to the second method even though both of the methods involve the multi-scale watershed segmentation of

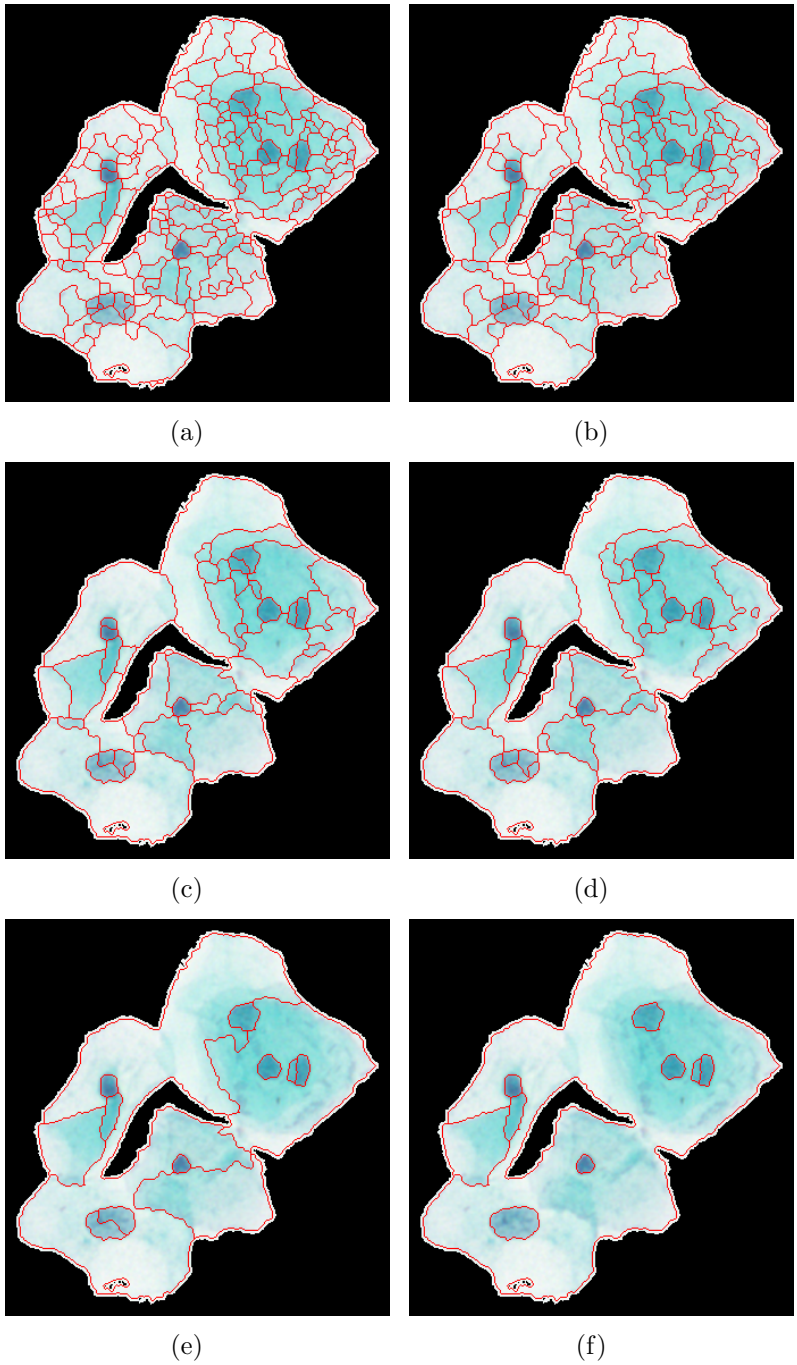


Figure 3.17: Candidate segments obtained by multi-scale watershed segmentation based on h-minima transform. (a) at scale 2 (b) at scale 3 (c) at scale 6 (d) at scale 7 (e) at scale 12 (f) at scale 13.

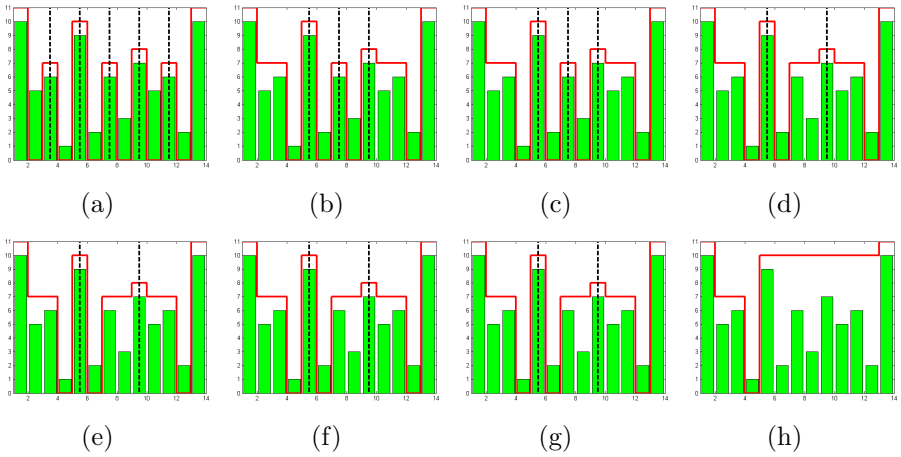


Figure 3.18: One-dimensional signal (green) and watersheds (black) of transformed signal (red) by minima imposition at scale (a) 0 (b) 1 (c) 2 (d) 3 (e) 4 (f) 5 (g) 6 (h) 7.

the image gradient by filtering the regional minima whose depth is lower than or equal to the corresponding threshold of each scale. The difference between these two methods stems from the technique used to filter the irrelevant regional minima of depth lower than or equal to the given threshold.

Figures 3.18 and 3.19 show the partitions of a one-dimensional input signal at different scales obtained using the second and third methods, respectively. The one-dimensional input signal is shown as the green colored bars. The red colored signal at each scale represents the transformation of the input signal resulting from elimination of the minima whose depth is lower than or equal to the corresponding threshold of each scale. Black dotted lines at each scale illustrate the corresponding partition determined as the watersheds of the transformed signal.

The minima imposition technique requires marker locations indicating relevant regional minima which are determined as the minima of depth higher than the corresponding threshold of each scale. The h-minima transform suppresses the irrelevant minima based on the parameter representing the maximum dynamic value of the minima to be filtered. The minima imposition technique does not alter the locations of relevant minima whereas the relevant minima locations

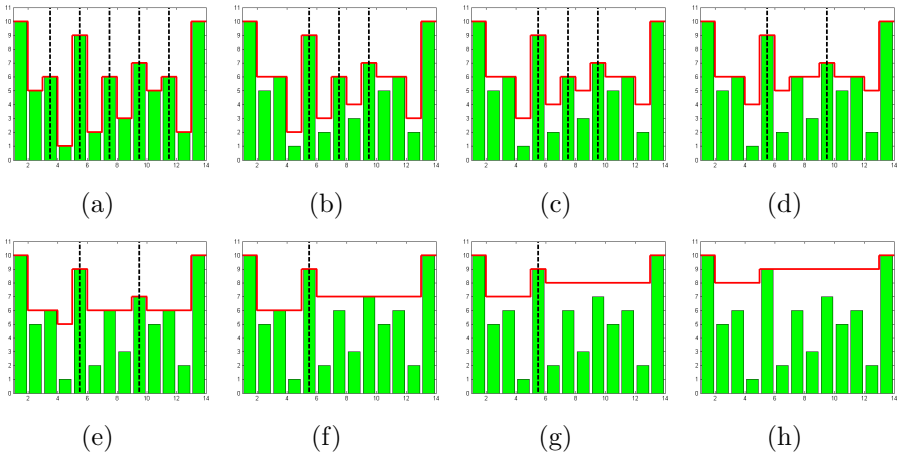
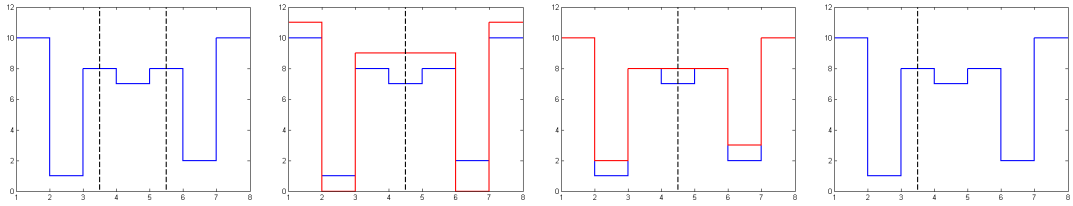


Figure 3.19: One-dimensional signal (green) and watersheds (black) of transformed signal (red) by h-minima at scale (a) 0 (b) 1 (c) 2 (d) 3 (e) 4 (f) 5 (g) 6 (h) 7.

either remain the same or are extended by the h-minima transform. For example, the minima locations of the signal transformed using two methods are different at scale 4 because some of the relevant minima are extended by the h-minima transform as can be seen from Figure 3.18 (e) and Figure 3.19 (e). The partition of the signal obtained by the h-minima transform differs from the corresponding partition of the signal obtained by the minima imposition technique when some of the relevant minima are extended by the h-minima transform so that they are merged with each other and contained in a single segment of the partition. This situation is illustrated in Figure 3.18 (f) and Figure 3.19 (f). The two minima of depth 7 are maintained at scale 5 by the minima imposition technique whereas these minima are merged by the h-minima transform at the same scale. Thus, we obtain different watershed partitions of the signal at scale 5 by two methods. Due to the same reason, the partitions at scale 6 also differ. This observation explains why two segments in Figure 3.13 constituting the upper nucleus at the partition of scale 13 obtained by the second method merge with cytoplasm region at the next scale without merging with each other whereas we obtain a true structure of the whole nucleus by the third method as shown in Figure 3.17.

A hierarchical tree can be constructed from the multi-scale partitions of a cell region if we ensure that the partitions are nested meaning that a region at a



(a) (b) (c) (d)

Figure 3.20: (a) Watershed partition of one-dimensional signal at scale 0 (b) partition of the second method at scale 1 (c) partition of the third method at scale 1 (d) each partition of scale 1 is adjusted.

partition of a certain scale either remains the same at the next scale or is contained in a bigger region by merging neighboring regions. The nested structure of the partitions obtained by the last two methods is disturbed when the image gradient has regional minima similar to the middle one in Figure 3.20 (a). The watershed partitions of the one-dimensional signal at scale 0 is same for the second and third method. Figure 3.20 (b) and (c) illustrate the watersheds of the signal transformed (red) at scale 1 by the second and third method, respectively. The middle region of scale 0 is split into two at the next scale for both methods because the watershed line between two regional minima of scale 1 is found at the center of the middle region of scale 0. We solve this problem by merging regions which are split in the next scale with their neighbor regions having the most similar mean intensity value. Mean intensity value of a region is defined as the average intensity of the pixels in that region. For example, if the middle region in Figure 3.20 (a) is more similar to the right region, we obtain the adjusted partition of scale 1 in Figure 3.20 (c) for both methods.

After comparing all of the methods for extracting hierarchical partitions of a cell region, we can conclude that the first method has the drawback of selecting a suitable size for the largest SE. The largest SE size should be large enough to generate candidate segments for all nuclei but when it gets larger, these segments merge with their surrounding noisy segments much. Moreover, both of the second and third methods involve the multi-scale watershed segmentation of the image gradient by filtering the irrelevant regional minima. The difference between the last two methods is that the minima imposition technique does not alter the

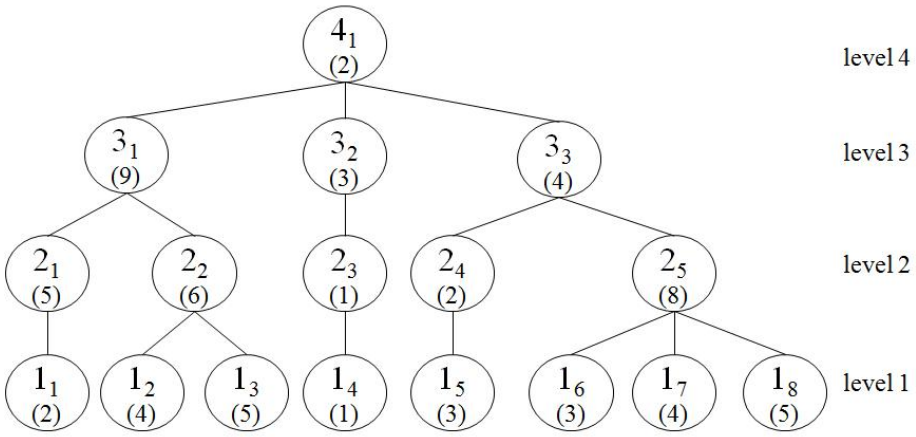


Figure 3.21: An example tree. Node i_j is a segment of the partition at scale i. j denotes the sequence number of the node from left to right in level i.

locations of relevant minima whereas the relevant minima locations either remain the same or are extended by the h-minima transform. Since we want the relevant marker locations to be extended as in Figure 3.17, the third method is selected for generating hierarchical partitions of a cell region.

We consider all regions of each partition as candidate meaningful segments. We construct a hierarchical tree from these regions where each segment is a node and there is an edge between two nodes corresponding to two consecutive scales if one node is contained within the other. Leaf nodes represent the regions extracted by watershed segmentation of the raw image gradient and the root node becomes the whole cell region. Figure 3.21 shows an example tree where the nodes are labeled as i_j with i denoting the node's level and j denotes the number of the node from left to right in level i. The node 3_1 has two children nodes 2_1 and 2_2, and its parent node is 4_1. The nodes 2_1 and 2_2 at level 2 combine to form the node 3_1 at level 3. The node 2_3 has only one child because the region represented as the node 1_4 does not merge with any other segments and remains the same at the next level.

3.2.2 Region Selection

In this section, our goal is to select the most meaningful segments among those appearing at different levels of the hierarchical tree described in the previous section. Each node of the tree is regarded as a candidate region in the final segmentation. Selection is done automatically as described in [1] but we calculate goodness measure for each node in a different way.

We consider the corresponding regions of the true nucleus structures as the most meaningful segments because the nucleus regions can be differentiated according to their homogeneity and shape properties. Besides, if the selected segments include all of the true nucleus structures, we can partition the cell region into segments of each nucleus and the whole cytoplasm area by classifying the selected segments as nucleus or cytoplasm and then forming the whole cytoplasm region as the union of the cytoplasm segments.

In general, small segments in the low levels of the hierarchy are merged to form larger segments of nucleus or cytoplasm area in the higher levels of the hierarchy. At some level, we obtain homogeneous and circular regions associated with true structures of nucleus regions. These nucleus regions may stay the same for some number of levels, and then, face a large change at a particular scale because they merge with their surrounding segments of cytoplasm area. Consequently, the segments we are interested in correspond to the homogeneous and circular regions right before this change. In other words, if the nodes on a path in the tree stay homogeneous and circular until some node n and then the homogeneity and circularity are lost in the next level, we say that n corresponds to a meaningful region in the hierarchy. Thus, the meaningfulness of a node is measured in terms of two factors: homogeneity and circularity. We compute the measure of each node starting from the leaf nodes up to the root node and select a node as a meaningful region if it is the most homogeneous and circular node on its path in the hierarchy where a path is defined as the set of nodes from a leaf to the root.

In order to calculate the circularity measure for a node in the hierarchy, we first find the eccentricity of the ellipse that has the same second-moments as the

corresponding region of the node. The eccentricity, e, is the ratio of the distance between the foci of the ellipse and its major axis length [39]. The value of the eccentricity is between 0 and 1 where an ellipse whose eccentricity is 0 is actually a circle, while an ellipse whose eccentricity is 1 is a line segment. We determine the circularity measure for the corresponding region of a node as $1/e$ such that more circular regions have higher values of the circularity measure. We add a very small value to the eccentricity for avoiding the division by zero problem that is encountered while calculating the circularity measure of a region whose shape is a circle. The circularity measure may favor small structures because in the extreme case a single pixel is considered as a circle. To overcome this problem, the circularity measure of the regions whose number of pixels is lower than a given threshold may be determined as 0.

As another factor, we introduce the homogeneity measure for a node based on the spectral features of the pixels in the corresponding region of the node. While examining a node from the leaf up to the root in terms of homogeneity, we consider the similarity between the node and its parent. We assume that the spectral features of the corresponding regions of the node R_1 and its parent R_2 having n_1 and n_2 pixels, respectively, are distributed normally. We first compute the mean and scatter of each region

$$\tilde{m_i} = \frac{1}{n_i} \sum_{y \in R_i} y \tag{3.13}$$

$$\tilde{s_i}^2 = \sum_{y \in R_i} (y - \tilde{m_i})^2 \tag{3.14}$$

by using maximum-likelihood estimation where y is a scalar value related to a pixel. If it can be assumed that the variances of the two regions are identical, then the statistic

$$F = \frac{(n_1 + n_2 - 2) n_1 n_2}{n_1 + n_2} \frac{(\tilde{m_1} - \tilde{m_2})^2}{\tilde{s_1}^2 + \tilde{s_2}^2} \tag{3.15}$$

has a small value under the hypothesis that the means of the regions are equal. For an F-value that is sufficiently large, the hypothesis can be rejected and the regions are found to be different from each other. We obtain a smaller F-value when the region remains the same in the next level or it is merged with similar regions to form its parent. When there is a large F-value, it means that the

region merged with the regions having different spectral features disturbing the homogeneity in the node. Thus, the F-value calculated between a node and its parent node should be maximized while selecting the most meaningful segments.

In order to compute the homogeneity factor of a node, the features of the corresponding pixels of the node and its parent should be one-dimensional. However, the pixels of each region in the hierarchy are represented by three-dimensional spectral features in the RGB color space. Thus, we project three-dimensional data consisting of the pixels of the node and its parent onto a line by using Fisher's linear discriminant analysis [14]. The goal of the discriminant analysis is to find an orientation for the line such that the projected data of the node and its parent are well separated.

We review the Fisher linear discriminant for completeness. Suppose that we have a set of three-dimensional pixels $\mathbf{x}$, n_1 of them in the corresponding region R_1 of the node and n_2 of them in the corresponding region R_2 of its parent node. We obtain the projection of $\mathbf{x}$ onto a line in the direction of $\mathbf{w}$ by the equation

$$y = \mathbf{w}^T \mathbf{x}. \tag{3.16}$$

The Fisher linear discriminant analysis employs $\mathbf{w}$ maximizing the criterion function

$$J(\mathbf{w}) = \frac{|\tilde{m_1} - \tilde{m_2}|^2}{\tilde{s_1}^2 + \tilde{s_2}^2} \tag{3.17}$$

which represents the measure of the good separation based on the difference between regions and the difference within the regions. To formulate the solution for $\mathbf{w}$, we calculate the mean for the data of each region as

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in R_i} \mathbf{x} \tag{3.18}$$

and the scatter matrices $\mathbf{S_i}$ and $\mathbf{S}_W$ as

$$\mathbf{S_i} = \sum_{\mathbf{x} \in R_i} (\mathbf{x} - \mathbf{m_i})(\mathbf{x} - \mathbf{m_i})^T \tag{3.19}$$

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2. \tag{3.20}$$

Then, the solution for $\mathbf{w}$ is derived in [14] as

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \tag{3.21}$$

The homogeneity measure for each node in the tree is calculated by the F-statistic between the projected data composed of the corresponding pixels of the node and its parent. As a result, the goodness measure M for a node n is defined as

$$M(n) = C(n) \times H(n, parent(n)) \tag{3.22}$$

where the first term is the node's circularity and the second term is the node's homogeneity measured according to its parent node. The node that is relatively homogeneous and circular will maximize this measure and will be selected as a meaningful region.

Given the value of the goodness measure for each node, the most meaningful regions are selected as described in [1]. A two-pass algorithm is used to select the most meaningful nodes in the tree. The first pass is bottom-up and its goal is to find the nodes each of which has a measure greater than all of its descendants. It first marks all of the nodes in level 1 and then starting from level 2 up to the root node, it continues to mark the nodes each of which has a measure greater than all of its descendants. Figure 3.22 shows the example run of the bottom-up algorithm for the example tree given in Figure 3.21 where the measure of each node is given in parenthesis.

The second pass is a top-down algorithm that starts by marking the root node as *selected* if it is already marked by the bottom-up algorithm. Then, the algorithm marks each node in each level until the leaf level as *selected* if it is marked in the first pass while none of its ancestors is marked. Finally, the nodes of the tree marked as *selected* are the resulting meaningful regions. Figure 3.23 shows the marked nodes in each step of the top-down algorithm for the example tree given in Figure 3.21.

Figure 3.24 illustrates the segmentation result of the example cell image where the boundaries of the obtained segments are colored with red. The final most meaningful segments include the true structures of all nucleus regions as well as other regions related to the cytoplasm area.

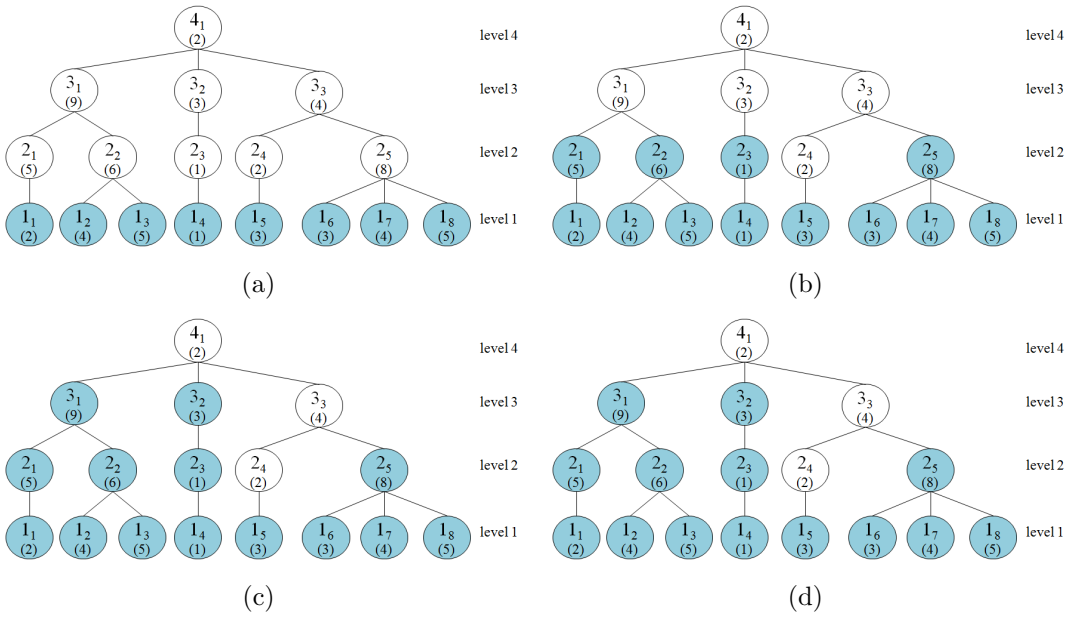


Figure 3.22: An example run of the bottom-up algorithm for the example tree given in Figure 3.21. Starting from level 1, the nodes having a measure greater than all of its descendants are colored with blue in each step.

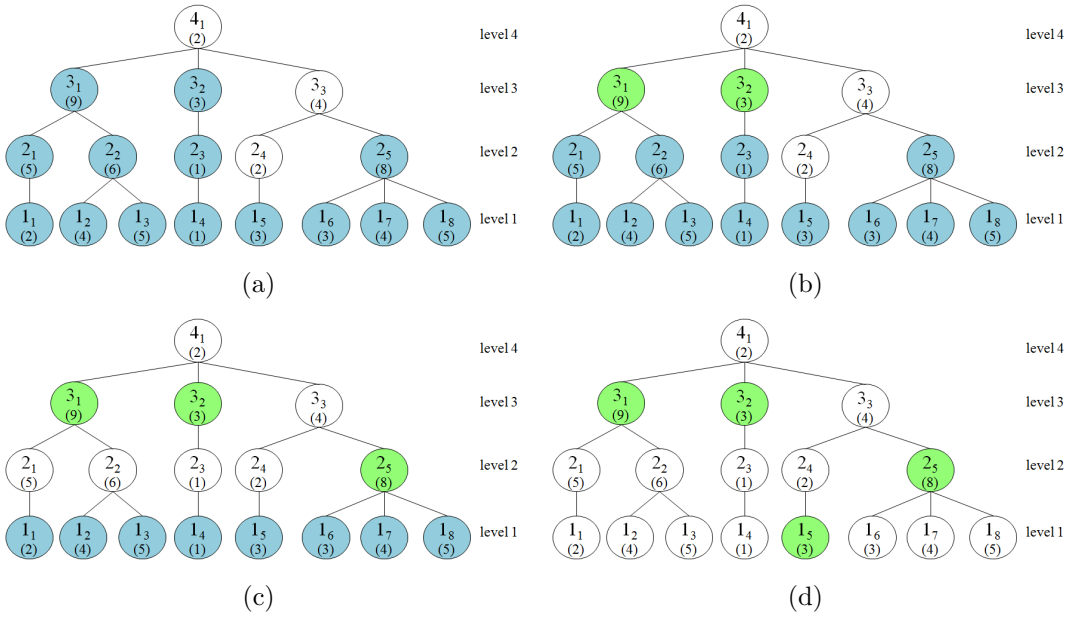


Figure 3.23: An example run of the top-down algorithm for the example tree given in Figure 3.21. Starting from the root level, the nodes marked in the first pass while none of its ancestors is marked are marked as *selected*. The green nodes are the final most meaningful segments.

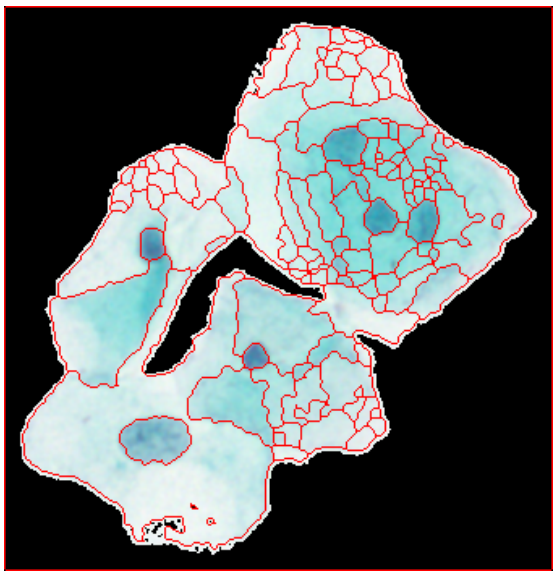


Figure 3.24: The segmentation result of the example cell image.

3.3 Nucleus and Cytoplasm Classification

In this phase, we aim to divide a cell region into true structures of each nucleus and the remaining whole cytoplasm area. We first classify each segment of the segmentation result obtained in the previous step as nucleus or cytoplasm area. Then, the whole cytoplasm area is determined as the union of all cytoplasm segments and the nucleus segments constitute true structures of each nucleus.

The classification of the segments is based on their four different features, namely, their size, mean intensity, eccentricity, and homogeneity measurements. The size is simply calculated as the number of pixels in the segment. The mean intensity feature is determined as the mean of intensity values of the pixels in the region. In order to obtain a single mean intensity value, we employ the illumination-corrected L channel used in the background extraction step. The last two features are composed of the eccentricity and homogeneity measurements computed in the previous step.

We form a data set composed of the corresponding feature vectors of 1452 nucleus regions and 7726 cytoplasm regions. While collecting data from a cell region, all of the nucleus and cytoplasm regions resulting from the segmentation

of that cell region are gathered in order to preserve the class frequencies.

The data set is partitioned into equally sized training and testing sets and the performances of different classifiers are experimented both on the original data and its normalized counterpart. Normalization is performed by linearly scaling each feature component to unit range where given the lower bound l and the upper bound u for a feature $x \in R$, the scaling $\widetilde{x} = \frac{x-l}{u-l}$ results in $\widetilde{x}$ being in the $[0, 1]$ range. Below, we describe the experimented classifiers and compare their performances.

3.3.1 Bayesian Classifier for Non-parametric Densities

Bayesian decision theory is a fundamental statistical approach that quantifies the tradeoffs between various decisions using probabilities and costs that accompany such decisions [14]. Our problem is a two-category classification problem where each segment belongs to either nucleus or cytoplasm area.

Let ω denote the state of nature, with $\omega = \omega_1$ for nucleus region and $\omega = \omega_2$ for cytoplasm region. Suppose also that action α_1 corresponds to deciding that the true state of nature is ω_1, and action α_2 corresponds to deciding that it is ω_2. The loss function $\lambda(\alpha_i|\omega_j)$ describes the loss incurred for taking action α_i when the state of nature is ω_j. Then, the conditional risks $R(\alpha_1|\mathbf{x})$ and $R(\alpha_2|\mathbf{x})$ associated with taking actions α_1 and α_2 given the four-dimensional feature vector $\mathbf{x}$ can be obtained as follows:

$$R(\alpha_1|\mathbf{x}) = \lambda(\alpha_1|\omega_1)P(\omega_1|\mathbf{x}) + \lambda(\alpha_1|\omega_2)P(\omega_2|\mathbf{x}) \tag{3.23}$$

$$R(\alpha_2|\mathbf{x}) = \lambda(\alpha_2|\omega_1)P(\omega_1|\mathbf{x}) + \lambda(\alpha_2|\omega_2)P(\omega_2|\mathbf{x}). \tag{3.24}$$

Minimum risk decision rule chooses ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$ and otherwise chooses ω_2. In terms of posterior probabilities, we decide ω_1 if

$$\left[\lambda(\alpha_2|\omega_1) - \lambda(\alpha_1|\omega_1)\right] P(\omega_1|\mathbf{x}) > \left[\lambda(\alpha_1|\omega_2) - \lambda(\alpha_2|\omega_2)\right] P(\omega_2|\mathbf{x}). \tag{3.25}$$

By using Bayes formula, we can replace the posterior probabilities by the prior probabilities and the conditional densities. This results in the equivalent rule, to

decide ω_1 if

$$[\lambda(\alpha_2|\omega_1) - \lambda(\alpha_1|\omega_1)]\, p(\mathbf{x}|\omega_1)P(\omega_1) > [\lambda(\alpha_1|\omega_2) - \lambda(\alpha_2|\omega_2)]\, p(\mathbf{x}|\omega_2)P(\omega_2) \tag{3.26}$$

and otherwise decide ω_2. The loss incurred for making an error is greater than the loss incurred for being correct so we can assume that both of the factors $[\lambda(\alpha_2|\omega_1) - \lambda(\alpha_1|\omega_1)]$ and $[\lambda(\alpha_1|\omega_2) - \lambda(\alpha_2|\omega_2)]$ are positive. An alternative formulation of the minimum risk decision rule can be obtained as follows:

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{[\lambda(\alpha_1|\omega_2) - \lambda(\alpha_2|\omega_2)]}{[\lambda(\alpha_2|\omega_1) - \lambda(\alpha_1|\omega_1)]} \frac{P(\omega_2)}{P(\omega_1)}. \tag{3.27}$$

In summary, the Bayesian decision rule can be interpreted as deciding ω_1 if the likelihood ratio $p(\mathbf{x}|\omega_1)/p(\mathbf{x}|\omega_2)$ exceeds a threshold value that is independent of the observation $\mathbf{x}$.

In order to calculate the likelihood ratio for a given feature vector $\mathbf{x}$, we assume that its four feature components are independent from each other. Then, the likelihood of ω_j with respect to $\mathbf{x} = [x_1\ x_2\ x_3\ x_4]$ can be computed as

$$p(\mathbf{x}|\omega_j) = p(x_1|\omega_j)p(x_2|\omega_j)p(x_3|\omega_j)p(x_4|\omega_j). \tag{3.28}$$

The class conditional density of each feature component is estimated by the simple histogram method. For each feature, we partition the space into a number of equally-sized bins and compute a histogram. Then the estimate of the density at a point x for a given class ω_j becomes

$$p(x|\omega_j) = \frac{k}{nV} \tag{3.29}$$

where n is the total number of samples in the class ω_j, k is the number of samples from the class ω_j in the cell that includes x, and V is the volume of that cell.

Figure 3.25 shows the class conditional probability density of each feature component estimated using the training data. We clip 5 and 1 percent of the right tail of the probability density functions associated with the size and the homogeneity features, respectively. By this way, we aim to eliminate outliers corresponding to the cytoplasm regions with very large size and the nucleus regions with very large F-value.

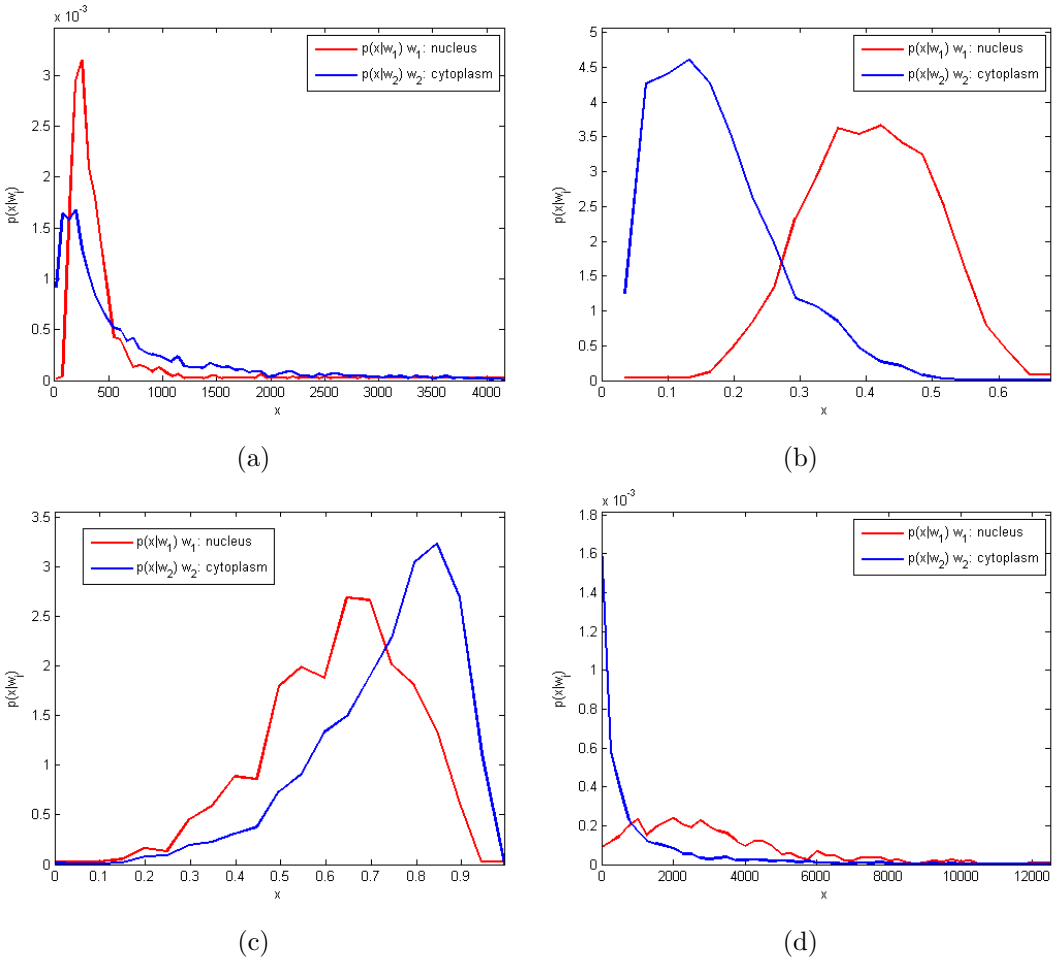


Figure 3.25: Class conditional probability density functions for the feature components (a) size (b) mean intensity (c) eccentricity (d) homogeneity.

After estimating the class conditional densities, the classification is performed for several meaningful combinations of the features and the results are investigated in order to find a suitable threshold for the likelihood ratio. Regarding each combination, the accurate classification rates are calculated for various thresholds by applying 10-fold cross validation on the whole data set (see Figure 3.26). We rerun the 10-fold cross validation 25 times to prevent the bias resulting from the partitioning of the data. The combination of all feature components gives higher classification performance for both nucleus and cytoplasm segments. Note that overall classification performance is governed by the cytoplasm segments because their number is much higher than the number of nucleus segments.

We want to find a threshold value for which the accurate classification rate is as high as possible for both nucleus and cytoplasm segments. The term $P(\omega_2)/P(\omega_1)$ on the right hand side of Equation (3.27) is estimated as 5.3 based on the class frequencies of the data set. The loss incurred for misclassifying nucleus regions $\lambda(\alpha_2|\omega_1)$ is assumed to be higher than the loss incurred for misclassifying cytoplasm segments $\lambda(\alpha_1|\omega_2)$ because we aim to obtain true structures of all nucleus segments. When the value for the threshold is determined as 2.6 under the assumption $\lambda(\alpha_2|\omega_1) = 2\,\lambda(\alpha_1|\omega_2)$, 51 nucleus regions and 196 cytoplasm regions in the testing data set are misclassified. The same classification results are obtained for the normalized features.

3.3.2 Bayesian Classifier for Normal Densities

Bayesian classification minimizing conditional risk becomes minimum error rate classification when we use the zero-one loss function which assigns no loss to a correct decision and assigns a unit loss to any error. Minimum error rate decision rule chooses ω_1 if $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$ and otherwise chooses ω_2.

One of the most useful ways to represent pattern classifiers is in terms of a set of discriminant functions $g_i(\mathbf{x}), i = 1, 2$. The classifier assigns a feature vector $\mathbf{x}$ to class ω_1 if $g_1(\mathbf{x}) > g_2(\mathbf{x})$. For minimum error rate case, we can let

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) \tag{3.30}$$

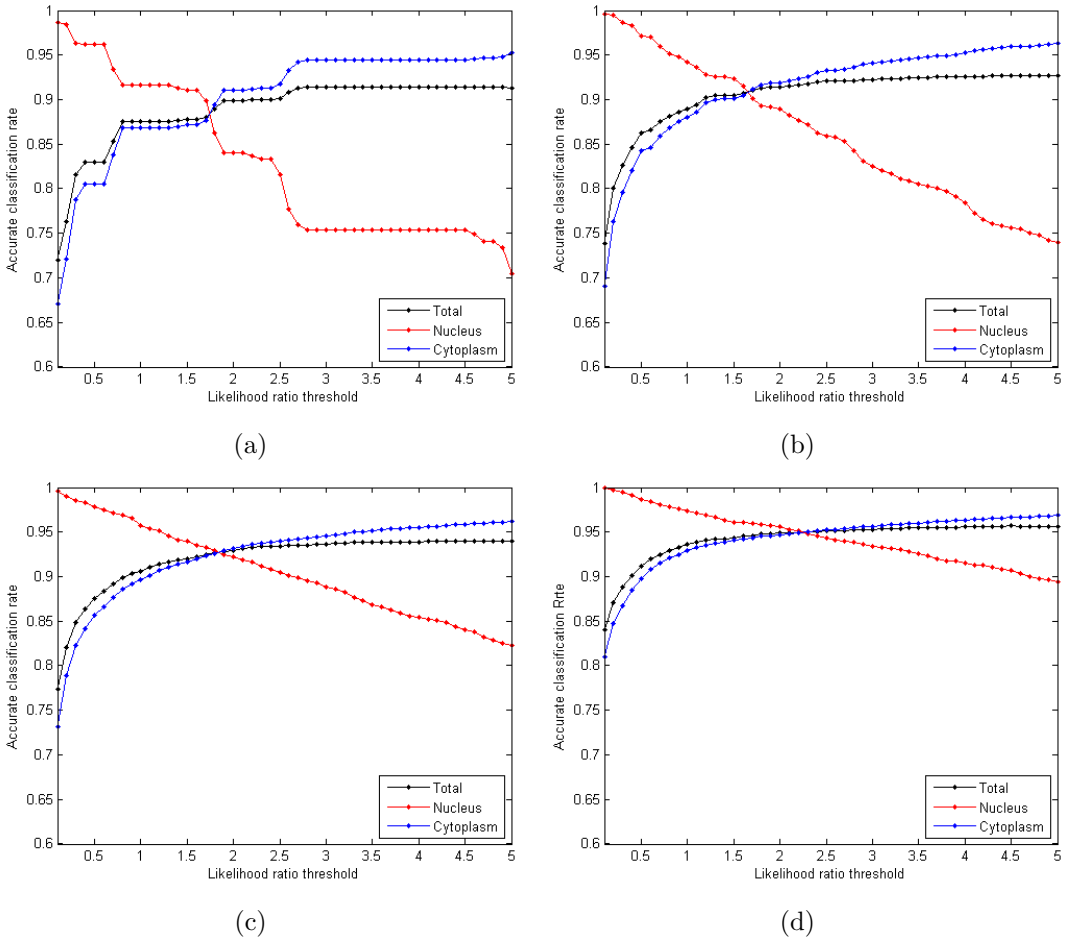


Figure 3.26: Likelihood ratio threshold versus accurate classification rate for nucleus (red) and cytoplasm (blue) segments given feature combinations (a) mean intensity (b) mean intensity and eccentricity (c) mean intensity, eccentricity and size (d) mean intensity, eccentricity, size and homogeneity.

because the maximum discriminant function will correspond to the maximum posterior probability. If we replace every $g_i(\mathbf{x})$ by $f(g_i(\mathbf{x}))$, where $f(.)$ is a monotonically increasing function, the resulting classification is unchanged. Then, the minimum error rate classification can be achieved by the use of discriminant functions

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i). \tag{3.31}$$

If we assume that the densities $p(\mathbf{x}|\omega_i)$ are multivariate normal that is $p(\mathbf{x}|\omega_i) \sim N(\mu_\mathbf{i}, \mathbf{\Sigma_i})$ then we have

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_\mathbf{i})^t \mathbf{\Sigma_i}^{-1}(\mathbf{x} - \mu_\mathbf{i}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\mathbf{\Sigma_i}| + \ln P(\omega_i). \tag{3.32}$$

We suppose that the covariance matrices are different for each class and the resulting discriminant functions becomes inherently quadratic [14].

The classifier is experimented on four different settings resulting from both determining prior probabilities as class frequencies or equal priors and normalizing features or not. The performance of the classifier for these four settings are summarized in Table 3.1. Using class frequencies as prior probabilities and normalizing features lower the error rate.

3.3.3 Decision Tree Classifier

Decision trees classify a pattern through a sequence of questions where the next question asked depends on the answer to the current question. This sort of classification is especially appropriate for nominal data which are discrete and without any natural notion of similarity or ordering. However, decision tree classifiers can also be used for numerical attributes like our features by means of questions that have the form whether the attribute x is less than x_0 or not.

In order to learn a decision tree, the set of training examples are partitioned into smaller and smaller subsets where each subset is as pure as possible. For a given node, the particular splitting attribute and the corresponding question is searched based on a purity measure. We can decide when to stop splitting using

thresholds on purity or the number of examples remaining at node, or statistical tests on the significance of reduction in impurity. Alternatively, the tree can be grown fully, and then can be pruned by considering the leaf nodes or even subtrees for elimination or merging.

For our classification problem, we compute a decision tree classifier out of the training set. We first expand the tree based on information gain as the binary splitting criterion and then apply pessimistic error pruning [31]. The splitting criterion and the pruning method is determined by performing cross-validation on the training set.

The performance of the decision tree on the testing data set is the same for both original and normalized features as illustrated in Table 3.1. The decision tree classifier results in lower misclassification rate when compared to many of the other classifiers experimented. This may be because decision trees do not have any assumptions about the distributions or the independence of the feature attributes. Furthermore, they automatically perform feature selection by using only the attributes partitioning the feature space more effectively.

3.3.4 Support Vector Classifier

Support Vector Classifier (SVC) is a popular technique used for data classification. Given a training set composed of instance-label pairs $(\mathbf{x}_i, y_i), i = 1, ..., n$ where $\mathbf{x}_i \in R^d$ and $y_i \in \{1, -1\}$, SVC requires the solution of the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ \textit{subject to} \quad & (\mathbf{w}^T.\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \quad \textit{for } y_i = +1, \\ & (\mathbf{w}^T.\phi(\mathbf{x}_i) + b) \leq -1 + \xi_i \quad \textit{for } y_i = -1, \\ & \xi_i \geq 0 \end{aligned} \tag{3.33}$$

Training data vectors $\mathbf{x}_i$ are mapped into a higher or infinite dimensional feature space by the function ϕ. Then, SVC finds a linear separating hyperplane with

the maximal margin in this higher dimensional space. Here, $C > 0$ is the penalty parameter of the error term such that lowering the value of C corresponds to a smaller penalty for the misclassification. $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$ is called the kernel function and various kernels are proposed in the literature [9].

For our classification problem, we choose radial basis function (RBF) kernel formulated as $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}$. The RBF kernel nonlinearly maps samples into a higher dimensional space such that it can handle the case when the relation between class labels and attributes is nonlinear. Moreover, the linear kernel is a special case of RBF [21].

There are two parameters C and γ while using RBF kernels. A grid search is performed on C and γ to find the best pair of parameters. Cross-validation is applied on the training data for each pair and the one with the best accuracy is selected. The whole training data along with the best parameters are used to train SVC and its classification performance is given in Table 3.1 for both original features and normalized features in the testing data.

3.3.5 Combined Classifiers

In this section, we combine different classifiers in order to improve the overall accuracy of the classification. The Bayesian classifier for normal densities using the class frequencies as the prior probabilities, the decision tree classifier and the SVC have similar error rates for the normalized testing data. However, the sets of instances misclassified by these classifiers do not necessarily overlap so we combine them in order to integrate their advantages. We prefer to use the second classifier in Table 3.1 instead of the first one because the misclassification rate of the nucleus regions is lower for the second one.

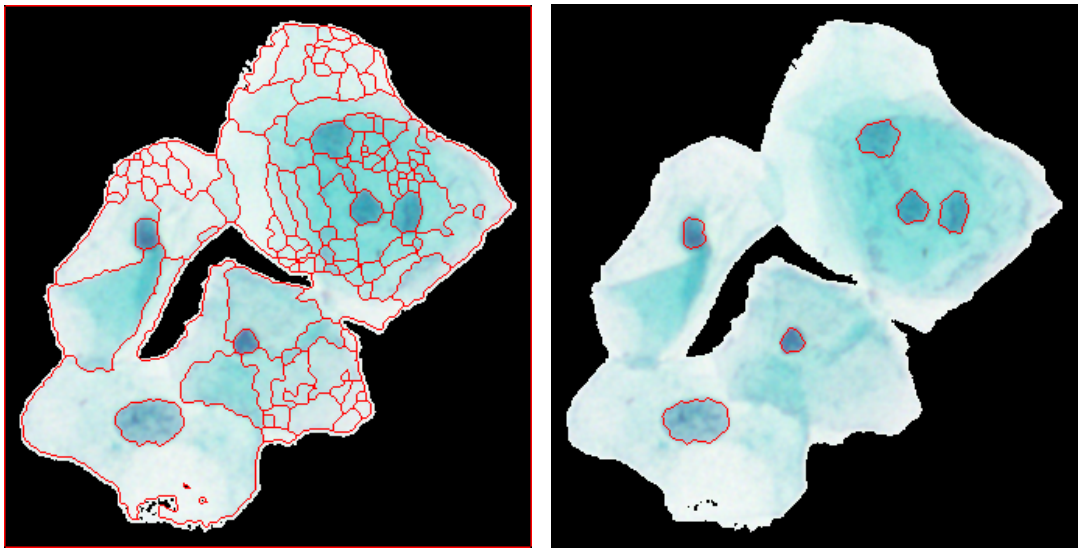
Classifiers are joined by three different combination schemes, namely majority voting, sum and product of the posterior probabilities computed by individual classifiers. In majority voting scheme, each classifier makes a vote for a single class and the decision is made in favor of the class with the largest number of votes. The

Table 3.1: Classification performances of different classifiers. The number of misclassified nucleus (N), cytoplasm (C) and total (T) regions in the testing data set are given for each classifier based on both original and normalized features.

		Original			**Normalized**		
	Classifier	**N**	**C**	**T**	**N**	**C**	**T**
1	Bayesian classifier Non-parametric densities	51	196	247	51	196	247
2	Bayesian classifier Normal densities, Class frequency	31	244	275	38	216	254
3	Bayesian classifier Normal densities, Equal priors	10	1460	1470	10	449	459
4	Decision tree	96	86	182	96	86	182
5	SVC with RBF	255	101	356	99	50	149
6	Combination of 2, 4 and 5 Majority voting	83	92	175	72	79	151
7	Combination of 2, 4 and 5 Sum	89	75	164	71	80	151
8	Combination of 2, 4 and 5 Product	108	52	160	65	96	161

performance of each combined classifier is presented in Table 3.1. Although the SVC on normalized data has the best performance in terms of the overall accuracy, we choose the combined classifier based on the sum of posterior probabilities. This is because it has higher accuracy for classification of the nucleus regions. We also prefer to normalize features which provides better classification results as can be observed from Table 3.1.

Figure 3.27 (b) shows the classification result for the example cell image where the boundaries of the nucleus segments are colored with red and the remaining area is determined as the whole cytoplasm. The segmentation result obtained for the image in the previous step is given in Figure 3.27 (a). We get true structures of each nucleus as well as the whole cytoplasm area. The whole cytoplasm area is found by taking union of all segments classified as cytoplasm.



(a) (b)

Figure 3.27: (a) The segmentation result (b) the classification result of the example cell image.

Chapter 4

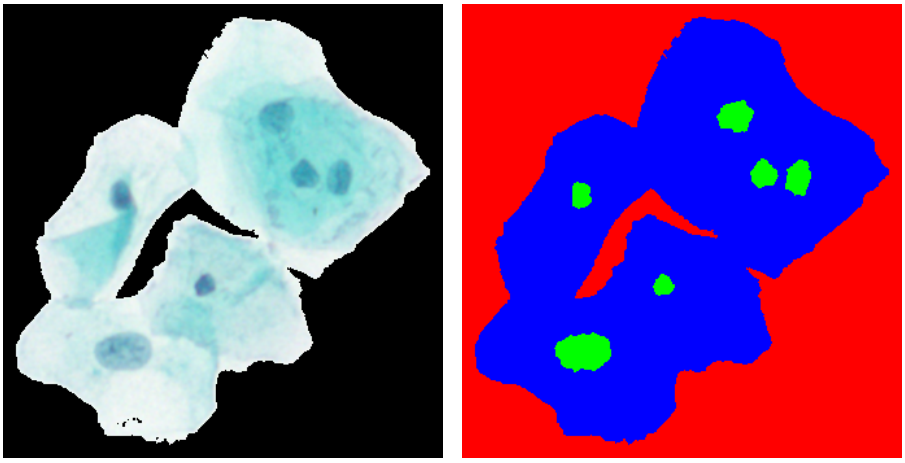
Classification of Cervical Cells

In this section, we first explain the features associated with cervical cells. Then, the procedure used to rank cervical cells is presented.

4.1 Feature Extraction

Dysplastic changes of cervical cells are associated with the cell characteristics like size, color, shape and texture of nucleus and cytoplasm. We describe each cell by using 14 different features related to these cell characteristics. The extracted features are a subset of the features used in [23] for characterizing cervical cells. We analyze the original and segmented cell images for feature extraction. In Figure 4.1, an example cell image and its segmentation and classification result are given where the red region represents background, the blue region corresponds to whole cytoplasm area, and the green regions represent the nucleus of each cell on the image.

In this step, our goal is to extract features for describing each cell in a cell image. A cell image may contain a single cell or a number of overlapping cells, and its segmentation and classification result consists of nucleus of each cell and a single cytoplasm area shared by all cells in the image. We associate each of



(a) (b)

Figure 4.1: (a) The example cell image, and (b) the corresponding segmentation and classification result.

the cells with their nucleus and the following features are extracted to obtain a description of them.

• *Nucleus Area*: Calculated by counting the number of pixels in the corresponding nucleus area of the segmented image.

• *Cytoplasm Area*: Calculated by dividing the total number of pixels in the whole cytoplasm area to the number of the cells in the image where the number of cells can be determined as the number of nucleus regions. We assume that the whole cytoplasm area is shared equally by all of the cells.

• *Nucleus/Cytoplasm Ratio*: This feature denotes how small the nucleus of a cell is when it is compared to the area of the cytoplasm associated to the cell. It is given by the ratio of the nucleus area to the cell area which is calculated as the sum of the nucleus and cytoplasm area.

• *Nucleus Brightness*: We calculate nucleus brightness as the average intensity of the pixels belonging to the nucleus region using the illumination-corrected L channel of the transformed image.

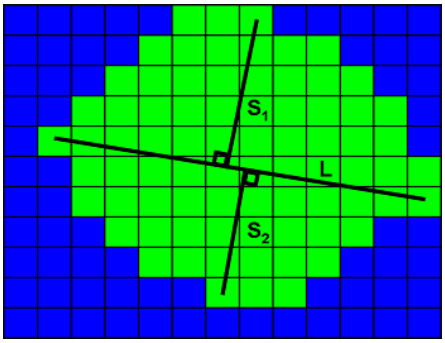


Figure 4.2: An example nucleus region (green) surrounded by cytoplasm region (blue). Longest diameter line L and shortest diameter lines S_1 and S_2 are shown.

• *Cytoplasm Brightness*: Cytoplasm brightness is obtained similar to the nucleus brightness. However, overlapping cells whose segmentation and classification results in a single cytoplasm region are associated with the same value of the cytoplasm brightness.

• *Nucleus Longest Diameter*: This feature corresponds to the diameter of the smallest circle circumscribing the nucleus region. We calculate it as the largest distance between the pixels on the border of the nucleus region. In Figure 4.2, the longest diameter L of an example nucleus region is illustrated.

• *Nucleus Shortest Diameter*: This is the diameter of the largest circle that is encircled by the nucleus region. It is approximated by the sum of the lines S_1 and S_2 which lie between the border pixels of the nucleus region and the line L. S_1 and S_2 are perpendicular to L, and they are the longest lines to each side of the region as shown in Figure 4.2.

• *Nucleus Elongation*: The nucleus elongation is calculated as the ratio between the shortest diameter and the longest diameter of the nucleus region.

• *Nucleus Roundness*: This feature is calculated as the ratio between the nucleus area and the area bounded the circle given by the nucleus longest diameter.

• *Nucleus Perimeter*: The length of the perimeter of the nucleus region.

• *Nucleus Maxima/Minima*: Calculated by counting the number of pixels each

of which is the maximum/minimum value inside of a 3x3 window centered on it.

• *Cytoplasm Maxima/Minima*: Calculated similar to the nucleus maxima and minima features. Overlapping cells whose segmentation and classification results in a single cytoplasm region are associated with the same value of these features.

4.2 Ranking of Cervical Cells

In this section, our purpose is to order cells in a Pap smear image according to their abnormality degree. In this way, the cells that are ranked as more normal than a selected cell that is manually identified as normal can be skipped by cyto-technicians or the cells can be investigated beginning from the end of the rank list that the most abnormal cells are found.

There are several reasons why we prefer to rank cervical cells instead of classifying them. First of all, as a supervised method, classification requires a large training set containing the complete repertoire of expected cell patterns for each class. Collecting such a training set is a very challenging task and entails a long period of time because cells on two slides may be quite different from each other due to artifacts, overlapping cells, and inconsistent staining. Furthermore, the complexities of cellular analysis and the need for high sensitivity and specificity make human intervention inevitable. Two semi-automated slide scanning devices approved by the FDA in the USA retrieve fields of diagnostic interest for examination of cyto-technicians rather than classifying slides [12]. Lastly, by proposing an unsupervised screening system, we aim to approach the problem in a different way when compared to the related studies that concentrates on classification.

In order to rank cells in a Pap slide, we first perform hierarchical clustering on the cell features extracted before. The initial ordering of the cells is determined as the leaf ordering of the constructed hierarchical tree. Then, this initial ordering is improved by applying the fast optimal leaf ordering algorithm [6].

Hierarchical clustering has been extensively used in biological literature to explore related genes that share a common function [6, 15, 34, 2, 8]. Vectors of gene expression levels are organized into a binary tree whose linear leaf ordering often discerns the underlying biological structure [6]. Hierarchical clustering produces a binary tree that groups a set of input elements over a variety of scales. The input elements are grouped according to their similarity which also determines their distance in the resulting binary tree.

To perform hierarchical clustering on a data set, we first compute the dissimilarity matrix between every pair of input elements using the Euclidean distance. After transforming the dissimilarity matrix into similarity matrix by subtracting each element from the maximum distance, we assign each element to a cluster. Then, the closest pair of clusters is merged at each step using the similarity matrix. After combining two clusters, we compute the similarity of the new cluster to the remaining ones. We repeat the last two steps until all elements are clustered into a single cluster. The key operation is the computation of the inter-cluster distances. The distance between two clusters is calculated using the average linkage method that uses the average distance between all pairs of elements in any two clusters.

The initial cell ordering is determined as the leaf ordering of the hierarchical tree. At each step, hierarchical clustering groups the closest pair of clusters so that the most related cells or cell groups become adjacent in the linear ordering of the tree. Figure 4.3 shows an example binary tree resulting from hierarchical clustering of 30 cells randomly selected from the Herlev data set. We select 5 cells from each class of normal superficial, normal intermediate, mild dysplasia, moderate dysplasia, severe dysplasia, and carcinoma in situ classes. As can be seen from the dendogram, the dysplastic cells are first organized into nested clusters, then the clusters of normal cells are formed. The clusters of dysplastic and normal cells are later merged into a single cluster.

Figure 4.5 gives the cell images and their class names in the linear leaf ordering of the tree shown in Figure 4.3. The dysplastic cells are found at the beginning of the ordering and the normal cells are grouped at the end of the list. However,

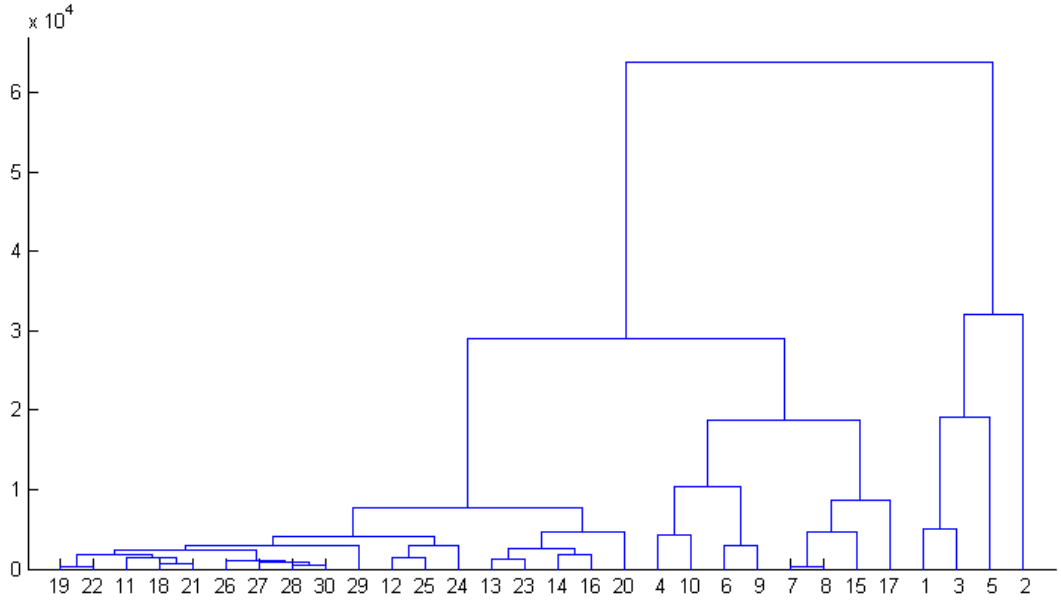


Figure 4.3: The binary tree resulting from hierarchical clustering of 30 cells randomly selected from the Herlev data. We select 5 cells from each class in order of normal superficial $(1-5)$, normal intermediate $(6-10)$, mild dysplasia $(11-15)$, moderate dysplasia $(16-20)$, severe dysplasia $(21-25)$, and carcinoma in situ $(26-30)$.

the dysplastic cells are not smoothly ordered according to their dysplasia degree and the group of normal cells at the end of the list contains some dysplastic cells. Hence, we need to improve the cell ordering obtained using hierarchical clustering.

In the related work of Bar-Joseph et al. [6], they present the first practical algorithm for the optimal linear leaf ordering of trees that are generated by hierarchical clustering. They present several examples showing that optimal leaf ordering is useful and achieves results that are superior to the heuristic ordering method presented in [15].

The optimal leaf ordering problem can be formulated as follows. Suppose that we have a binary tree T with n leaves $z_1, ..., z_n$ and $n-1$ internal nodes. We can obtain a linear ordering consistent with T by flipping the internal nodes in T. For example, as shown in Figure 4.4, when we flip the two subtrees rooted at the

(a) (b)

Figure 4.4: (a) An example binary tree T and (b) a linear leaf ordering consistent with T obtained by flipping the node marked by red ellipse.

node marked by the red ellipse, the ordering of the leaves is changed while the same tree structure is maintained. There are 2^{n-1} possible linear leaf orderings of the tree because it has $n-1$ internal nodes. The optimal leaf ordering problem is to find an ordering of the tree leaves that maximizes an optimization criterion. There are two types of optimization criteria which are described below.

The first optimization criterion maximizes the sum of similarities between adjacent leaves in the ordering. If we denote the space of the 2^{n-1} possible ordering of the tree leaves by Φ then for $\phi \in \Phi$, $D^{\phi}(T)$ is defined as

$$D^{\phi}(T) = \sum_{i=1}^{n-1} S(z_{\phi_i}, z_{\phi_{i+1}}) \tag{4.1}$$

where z_{ϕ_i} is the ith leaf when T is ordered according to ϕ and S is the similarity matrix. Thus, the first optimization criterion aims to find an ordering ϕ maximizing $D^{\phi}(T)$.

The second optimization criterion maximizes the sum of similarities between every leaf and all other leaves in the adjacent clusters. For the tree shown in Figure 4.4 (a), the left adjacent cluster of the leaf 3 contains the leaf 1 and 2 whereas its right adjacent cluster consists of the leaf 4 and 5. Then, the set of the similarities between the leaf 3 and its adjacent clusters becomes $\{S(3,1), S(3,2), S(3,4), S(3,5)\}$. As a result, the criterion value for this tree can be calculated as the sum of the elements in the union of the sets $\{S(1,2)\}$, $\{S(2,1), S(2,3), S(2,4), S(2,5)\}$, $\{S(3,1), S(3,2), S(3,4), S(3,5)\}$, $\{S(4,3), S(4,5)\}$, and $\{S(5,4)\}$.

The optimal leaf ordering algorithm runs in $O(n^4)$ time and $O(n^2)$ space. The

algorithm is further improved to make its running time very reasonable when compared to the hierarchical clustering which is implemented to run in $O(n^3)$ time. More details can be found in [6].

Cells in a Pap smear slide are ranked by applying the optimal leaf ordering algorithm to the binary tree obtained by hierarchical clustering. The second optimization criterion is chosen for the optimal leaf ordering algorithm because it considers the similarities between a leaf and its adjacent clusters rather than its adjacent leaves which makes the ordering less sensitive to noise and outliers. An example cell ranking obtained by the optimal leaf ordering algorithm is given in Figure 4.6 from which we can observe that the initial ordering in Figure 4.5 is improved.

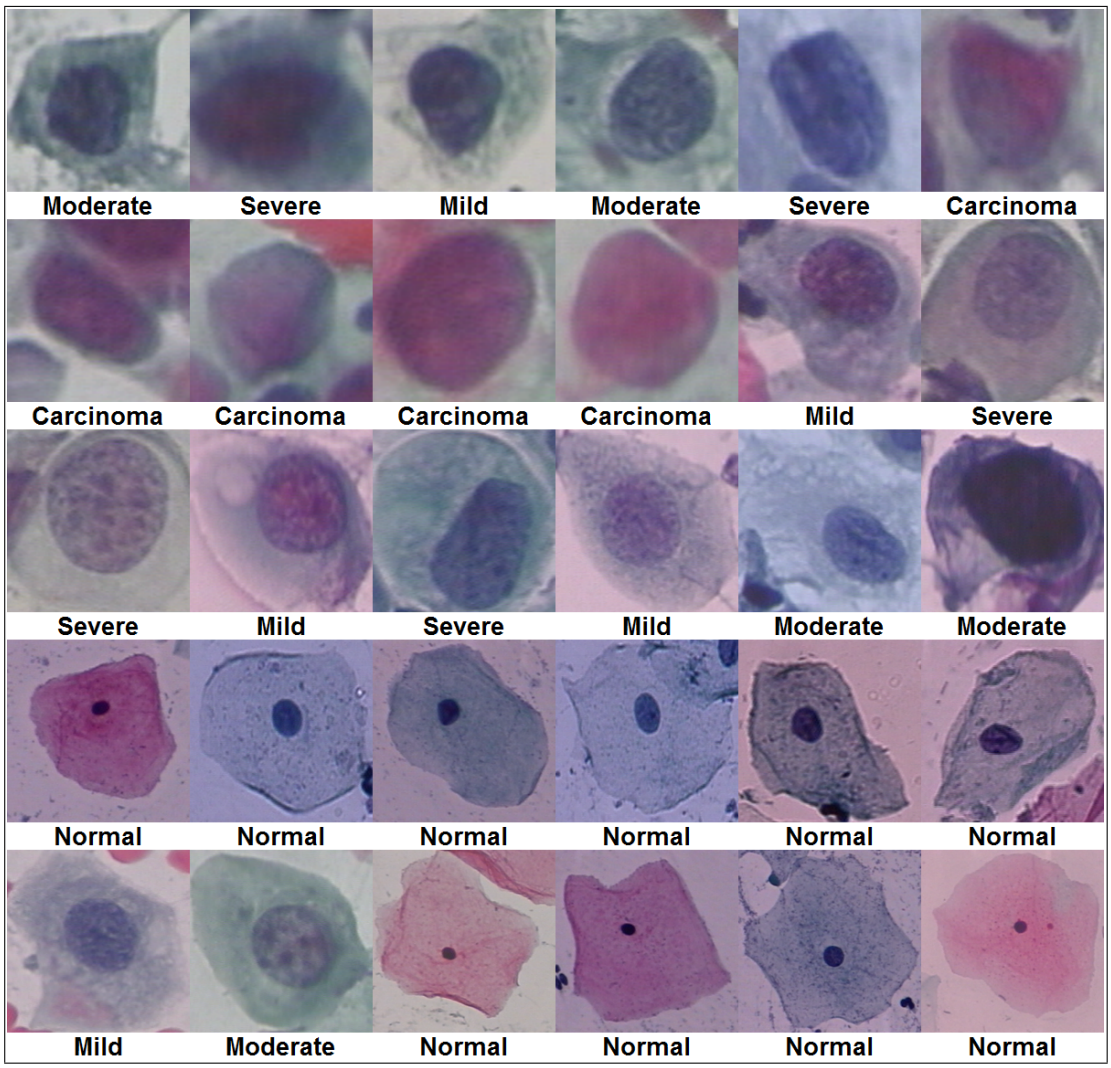


Figure 4.5: The initial ordering of the cells determined as the linear ordering of the leaves of the tree in Figure 4.3. (The cell images are resized to have the same width and height so their relative size is not proper.)

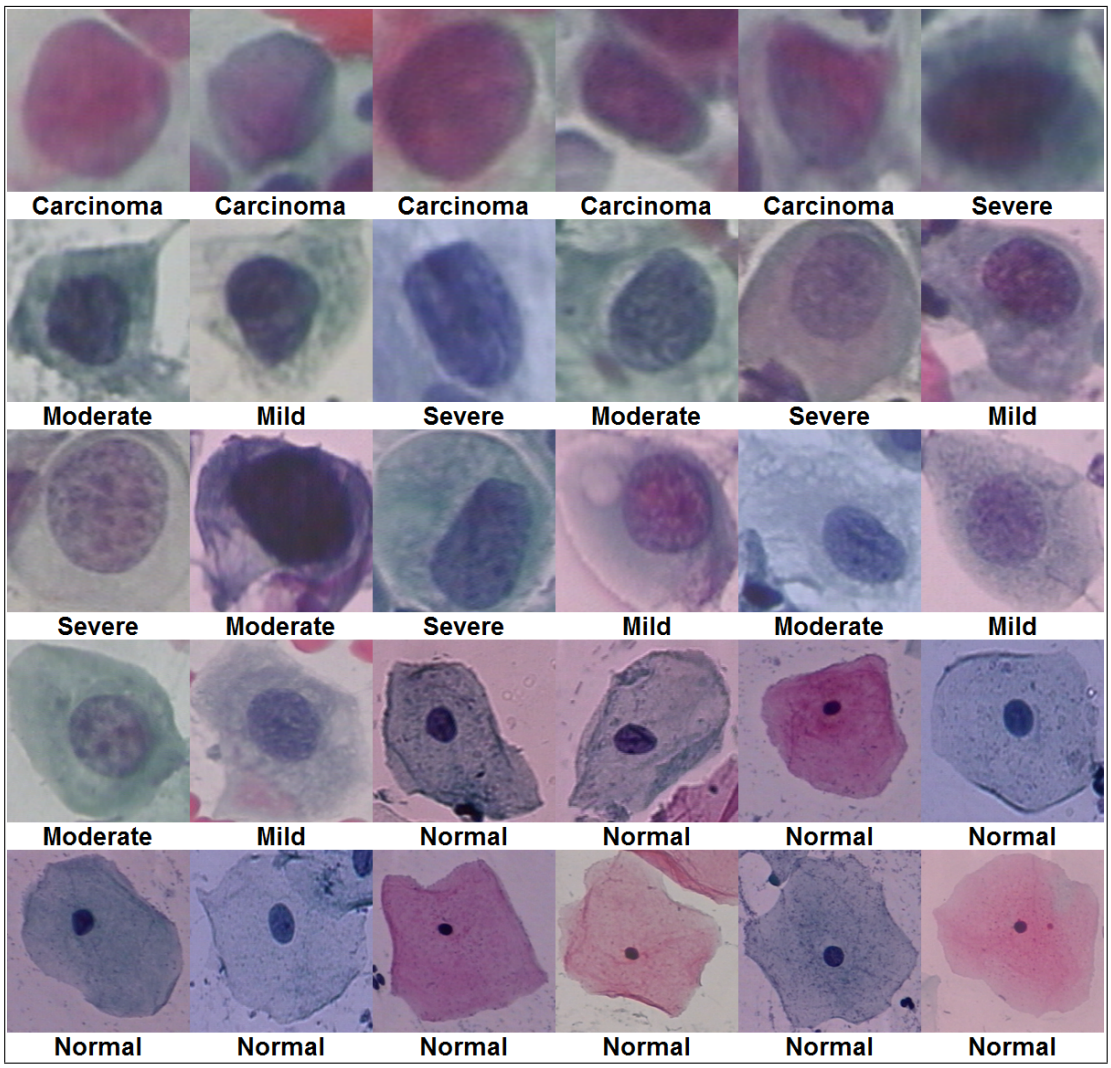


Figure 4.6: The final ordering of the cells obtained by applying the optimal leaf ordering algorithm to the initial ordering in Figure 4.5. (The cell images are resized to have the same width and height so their relative size is not proper.)

Chapter 5

Experiments and Results

In this part, we present the experiments conducted for the proposed work and discuss their results. The experiments are performed on the Herlev data and the Hacettepe data.

5.1 Segmentation of Cervical Cells

In this thesis, we propose a three-phase approach to segmentation of cervical cells where the first phase involves background extraction for obtaining remaining cell regions. The second phase consists of segmenting the cell regions by a non-parametric hierarchical segmentation algorithm. In the last phase, we classify the segments of the cell region as nucleus or cytoplasm in order to determine true structures of each nucleus and the remaining whole cytoplasm area.

The experimental results obtained for each phase of the segmentation algorithm are presented below. We first evaluate the background extraction on the Hacettepe data and give some illustrative results. Then, the performance of our non-parametric hierarchical segmentation algorithm for locating nucleus regions along with their boundaries is compared against the ground truth of the Herlev data using the Zijdenbos similarity index (ZSI) [45]. We also discuss the behavior

of the algorithm in the case of complicated cell regions by using representative examples. Lastly, the classification accuracy for identifying nucleus and cytoplasm regions is presented using the Hacettepe data.

5.1.1 Background Extraction

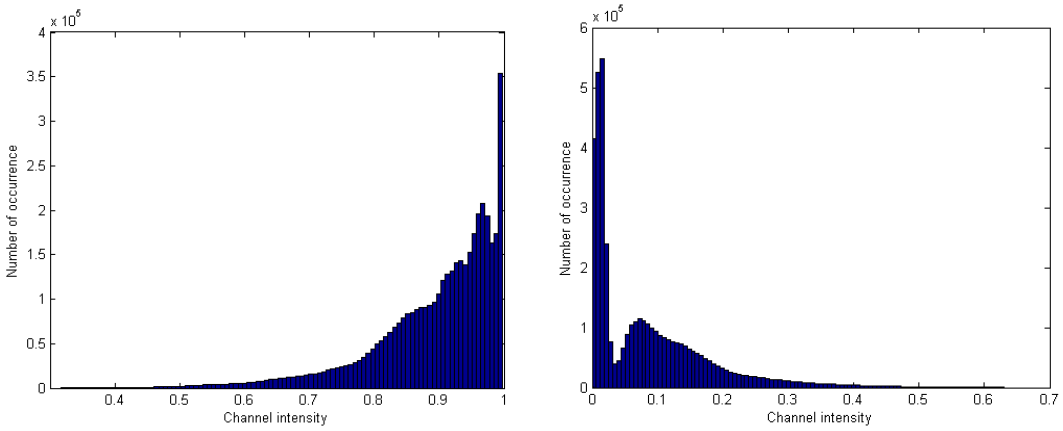
It is not necessary to perform background extraction on the Herlev data since it consists of single cell images many of which do not have any background areas. On the other hand, there is no ground truth data involving the boundaries between the cell regions and the background area in the images of the Hacettepe data. Hence, in Figure 5.1 to 5.3, we give illustrative examples covering a wide range of the images existing in the Hacettepe data to evaluate the performance of this phase.

In each example, the histogram of the L channel obtained by transforming the image from the RGB color space to the CIE Lab color space is given in (a), the histogram of the illumination-corrected L channel is given in (b) and the boundaries of the segmented cell regions obtained by this phase are given in (c).

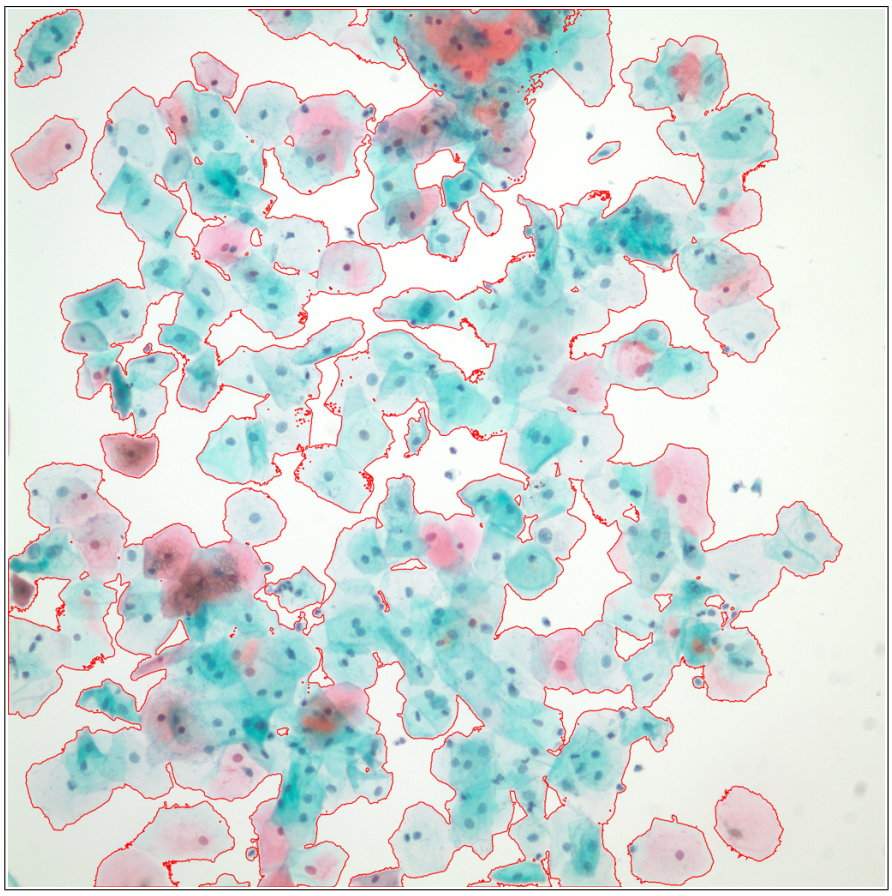
The example image shown in Figure 5.1 (c) comprises a number of cell regions with so many overlapping cells. After filtering non-homogeneous illumination, we obtain the bimodal histogram in (b). We can observe from (c) that the background is smoothly extracted for a Pap test image consisting of many cells.

Figure 5.2 illustrates an example image with a limited number of cells. The cell regions are extracted very well, although two of them are false detections. These false cell regions are resulted due to the intensive non-homogeneous illumination of the Pap smear slide. The false cell region with an oval shape in the middle of the image may be followed by the pollution in this part of the slide.

As opposed to the previous images, the cell density in the image of Figure 5.3 is in between the corresponding cell densities of the first and second images. The cell regions in this image are colored with different color tones compared to the previous images due to the inconsistent staining. The background extraction is

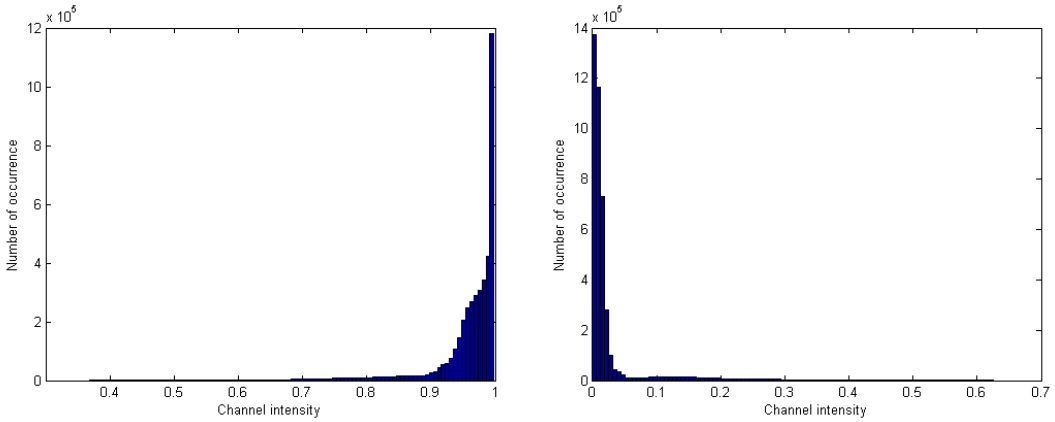


(a) (b)

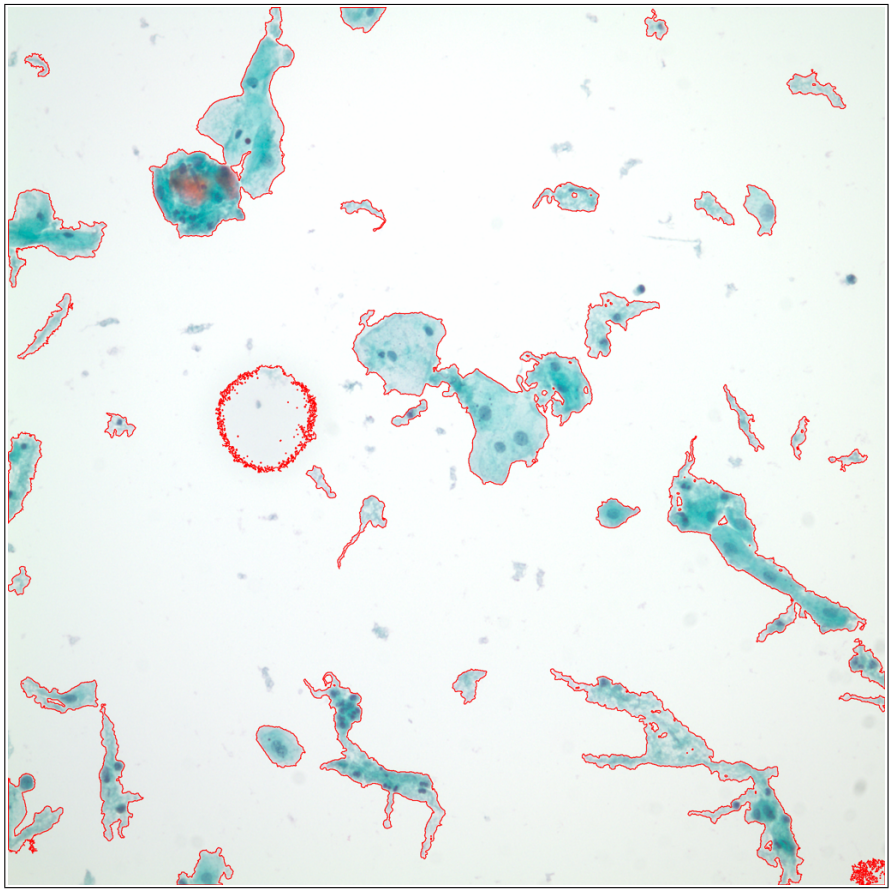


(c)

Figure 5.1: (a) The histogram of the L channel (b) the histogram of the illumination-corrected L channel (c) the boundaries of the segmented cell regions are colored as red.

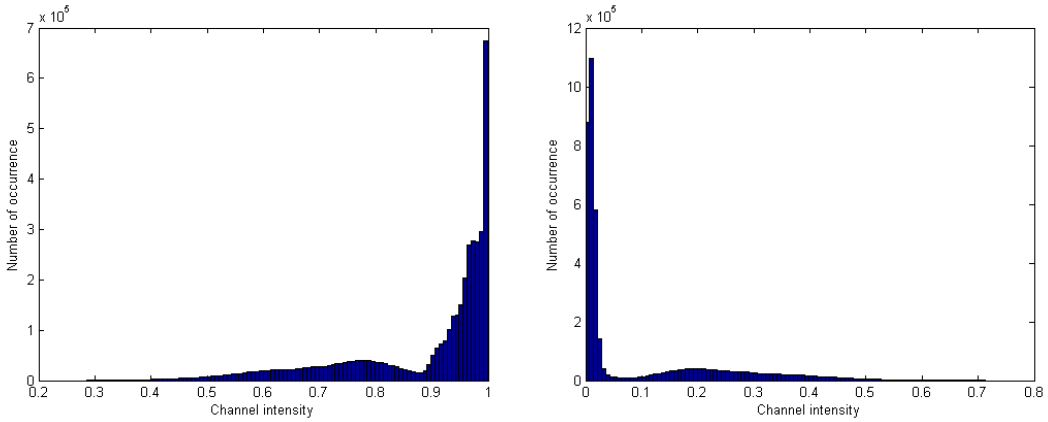


(a) (b)

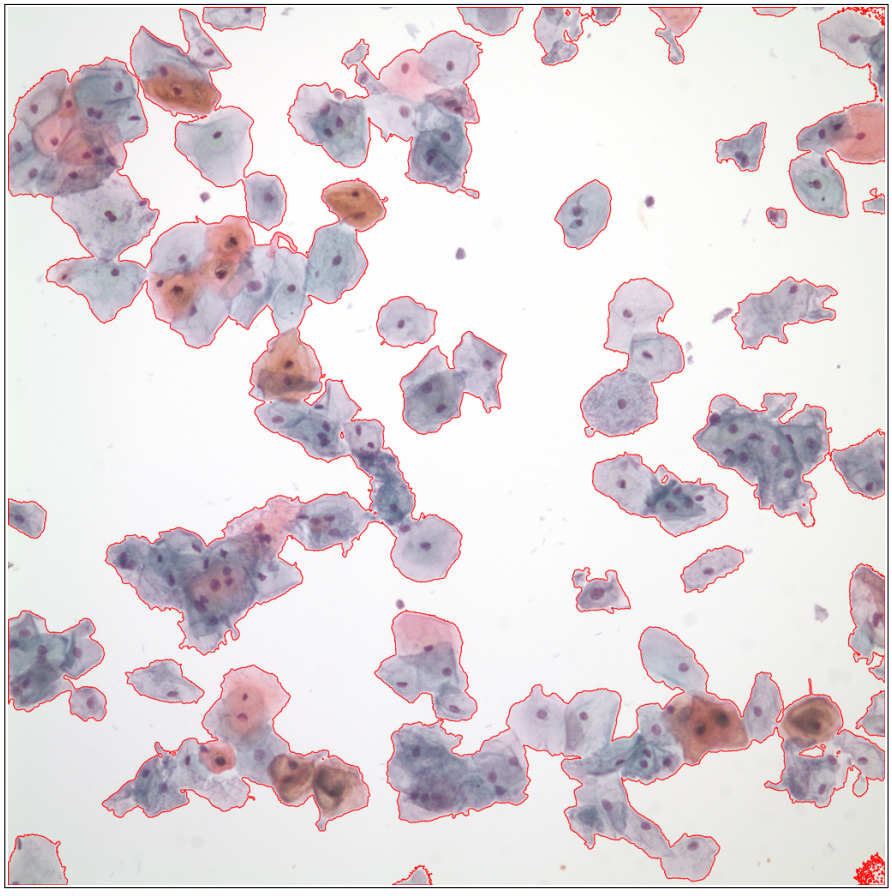


(c)

Figure 5.2: (a) The histogram of the L channel (b) the histogram of the illumination-corrected L channel (c) the boundaries of the segmented cell regions are colored as red.



(a) (b)



(c)

Figure 5.3: (a) The histogram of the L channel (b) the histogram of the illumination-corrected L channel (c) the boundaries of the segmented cell regions are colored as red.

performed well on this image except the false cell region detected in the bottom-right corner.

To sum up, the background extraction phase performs well under varying conditions and it is a very practical method based on simple thresholding. It only suffers from the intensive non-homogeneous illumination especially on the image corners. The uneven illumination of the images arises from the image acquisition stage which can be improved using a better controlled setup to overcome this problem.

5.1.2 Nucleus and Cytoplasm Segmentation

Our non-parametric hierarchical segmentation algorithm aims to find a partitioning of a cell region where true structure of each nucleus is captured well by its corresponding segment. In order to evaluate our segmentation method, we first perform its quantitative analysis by using the ground truth of the Herlev data. The Herlev data includes correct segmentation results for single cell images of all classes. The performance of our segmentation algorithm for locating nucleus regions along with their boundaries is compared against the ground truth of the Herlev data using the Zijdenbos similarity index (ZSI) [45].

The ZSI is defined as the ratio of twice the common area between two regions A_1 and A_2 to the sum of the individual areas as $S = 2\,\frac{|A_1 \cap A_2|}{|A_1|+|A_2|}$ where $S \in [0, 1]$. This similarity index is sensitive to differences in both size and location [45]. However, the differences in location are more strongly reflected than the differences in size. For example, the similarity of two equally sized regions that overlap each other with the half of their area is $1/2$ whereas the similarity of two regions one of which completely overlapping the smaller one whose size is half of the bigger one is calculated as $2/3$. This follows the intuition that two regions one of which fully contains the other are more similar than two partially overlapping regions.

Given a cell image from the Herlev data, the performance of our segmentation

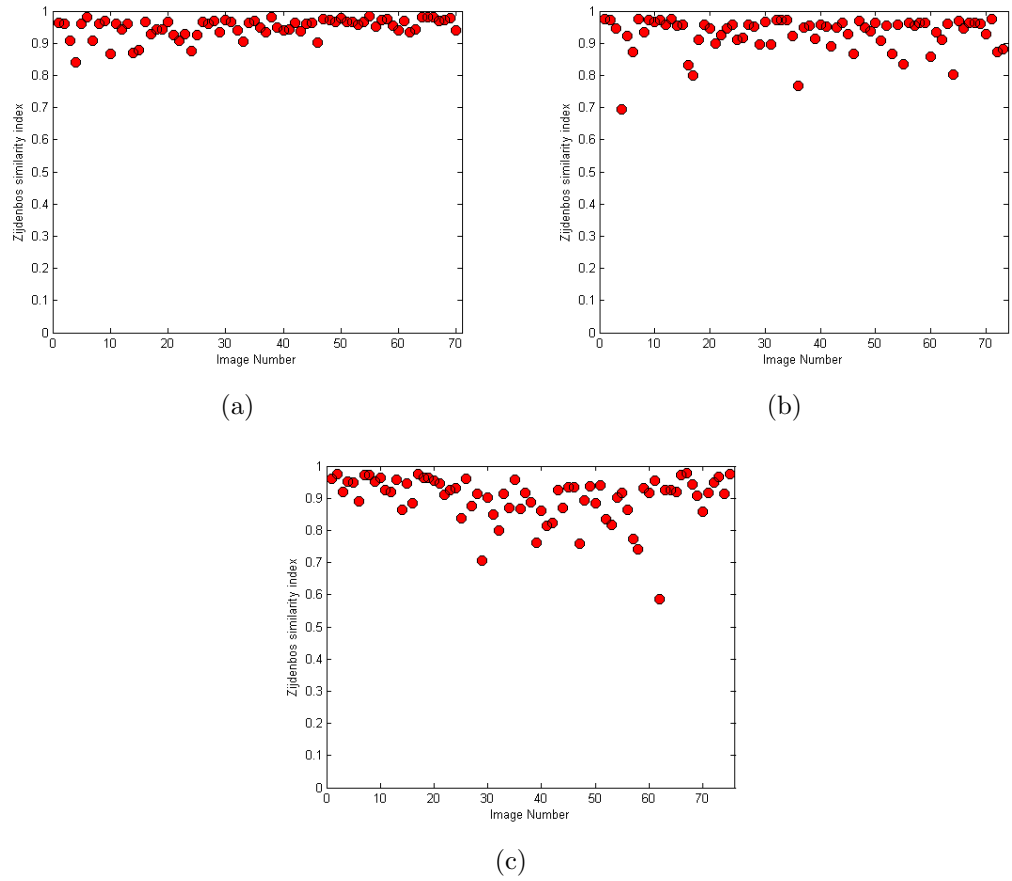


Figure 5.4: The ZSIs for the images of the classes (a) Intermediate squamous (b) Superficial squamous (c) Columnar

method for locating the nucleus region along with its boundary is measured as follows: After segmenting the cell image by our method, we compute the ZSIs for the ground truth of true nucleus region compared with each of the resulting segments overlapping this nucleus region. Since the segment with the highest ZSI is more likely to be the corresponding segment of the true nucleus structure, the performance of our method for the given cell image is measured by this highest ZSI knowing that the value of ZSI greater than 0.7 indicates excellent agreement between the segments [45]. Figures 5.4 and 5.5 show the calculated ZSIs for the images of all classes and the ZSI means and standard deviations for each class are given in Table 5.1. The ZSI for our segmentation of the nucleus region compared with the ground truth segmentation has a mean larger than 0.90 and a standard deviation smaller than 0.2 for all classes. Overall, our segmentation method

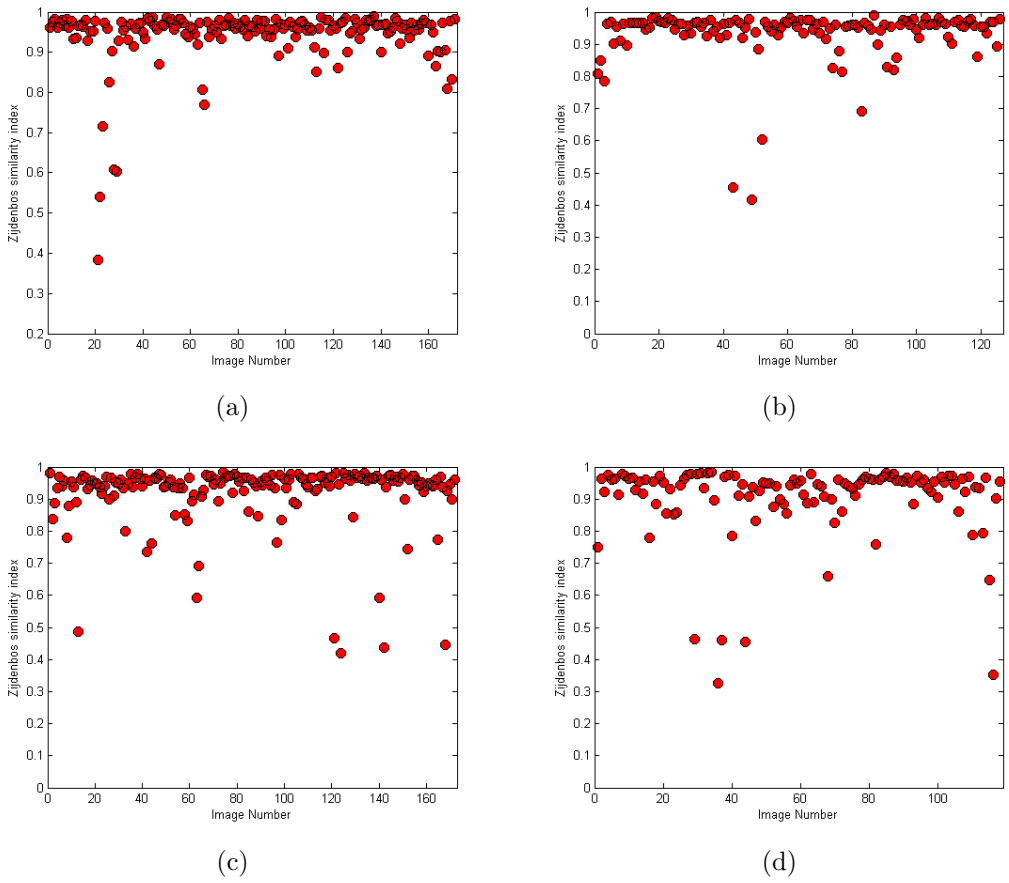


Figure 5.5: The ZSIs for the images of the classes (a) Mild dysplasia (b) Moderate dysplasia (c) Severe dysplasia (d) Carcinoma in situ

Table 5.1: The ZSI means and standard deviations of each class for the ground truth compared to our segmentation.

	ZSI mean $\pm$ standard deviation
Superficiel squamous	0.93 ± 0.05
Intermediate squamous	0.95 ± 0.03
Columnar	0.90 ± 0.07
Mild dysplasia	0.94 ± 0.08
Moderate dysplasia	0.93 ± 0.08
Severe dysplasia	0.92 ± 0.10
Carcinoma in situ	0.90 ± 0.12

performs well except some cell images whose nucleus region cannot be segmented correctly due to the reasons that will be clarified further in this section.

Different from the related studies in the literature [44, 40], we propose a generic segmentation algorithm that can be applied to the cell images containing overlapping cells. Since there is no ground truth involving correct segmentation results of the images in the Hacettepe data, we evaluate our method on the Hacettepe data qualitatively by giving representative examples.

Segmentation results of example images from the Hacettepe data are given in Figures 5.6 to 5.10. In each sample, the image and the result of our segmentation method are shown. Our goal of segmenting an image of a cell region is to obtain a corresponding segment for each nucleus that captures the true structure of that nucleus well. Hence, it does not matter how the cytoplasm area of the cell region is partitioned into many segments because the obtained segments are later to be classified as nucleus or cytoplasm based on their features.

We obtain the true structures of nucleus regions in the images of Figures 5.6 to 5.10 using our segmentation method. This means that for each image, we obtain these nucleus regions at some levels of the hierarchical tree constructed by our algorithm. Moreover, the segments associated with these nucleus regions are the most meaningful nodes on their paths in the hierarchy. The meaningfulness of a region is measured in terms of its homogeneity and circularity but our method is

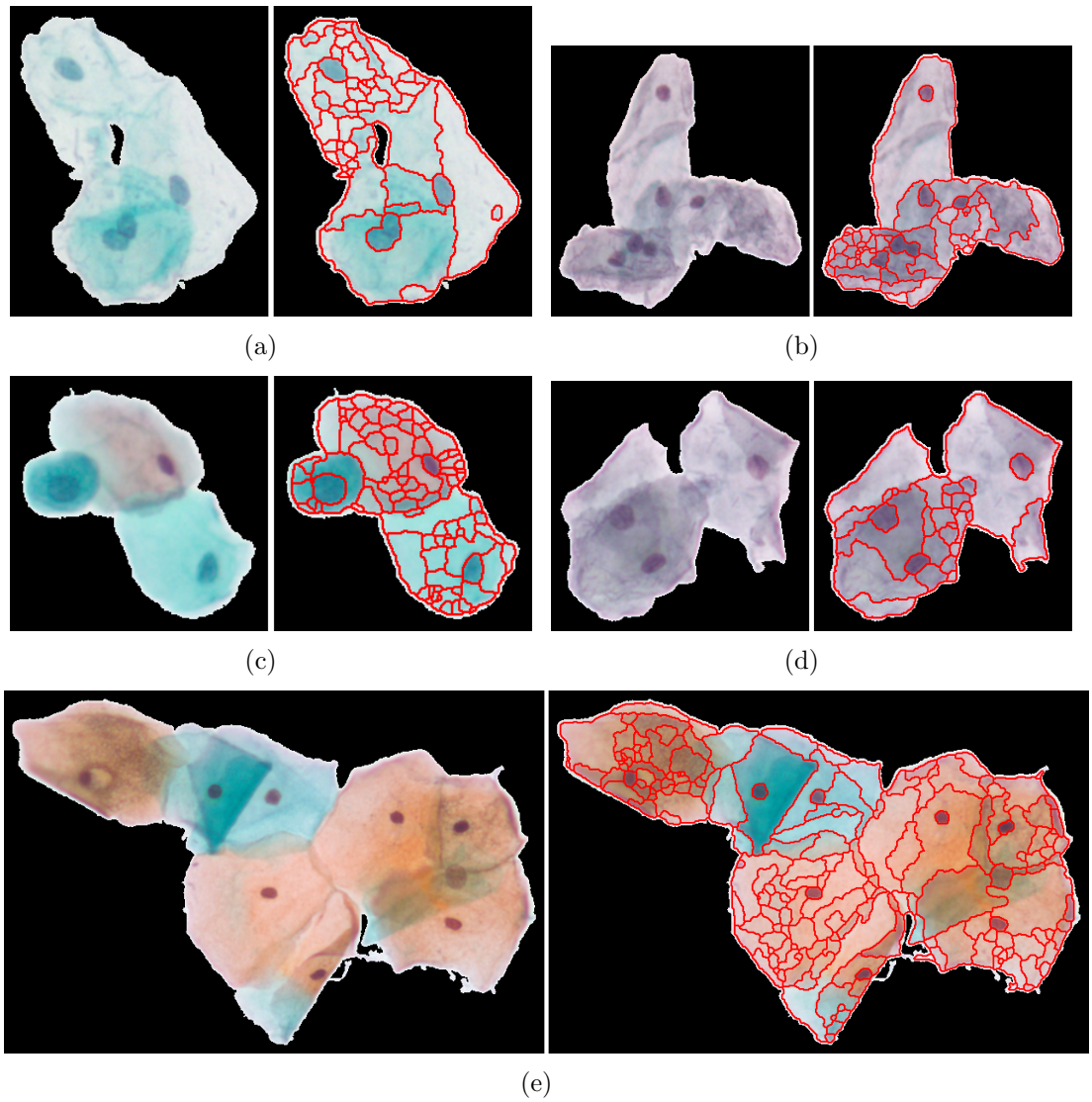
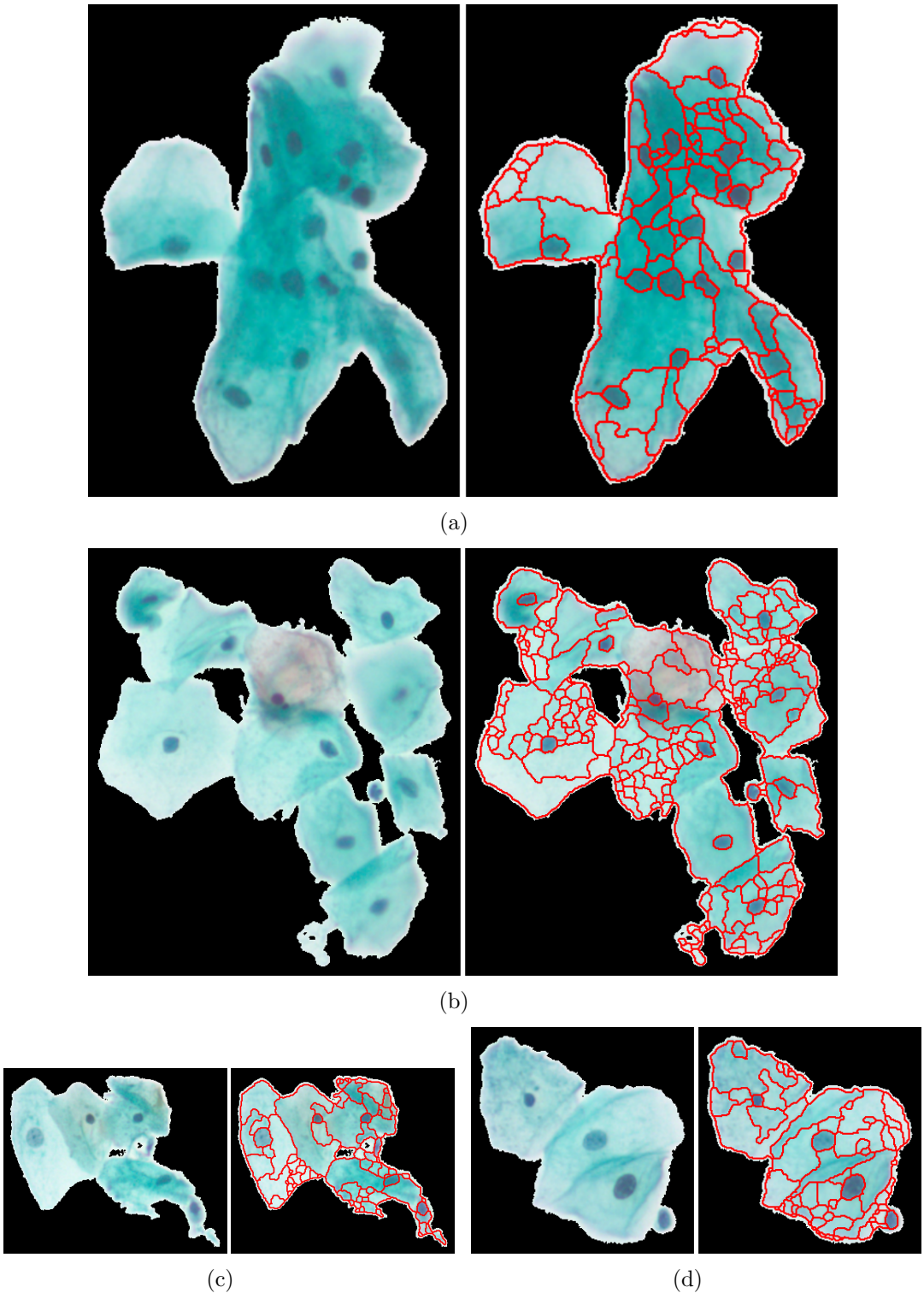


Figure 5.6: Segmentation results for example images from Hacettepe data.

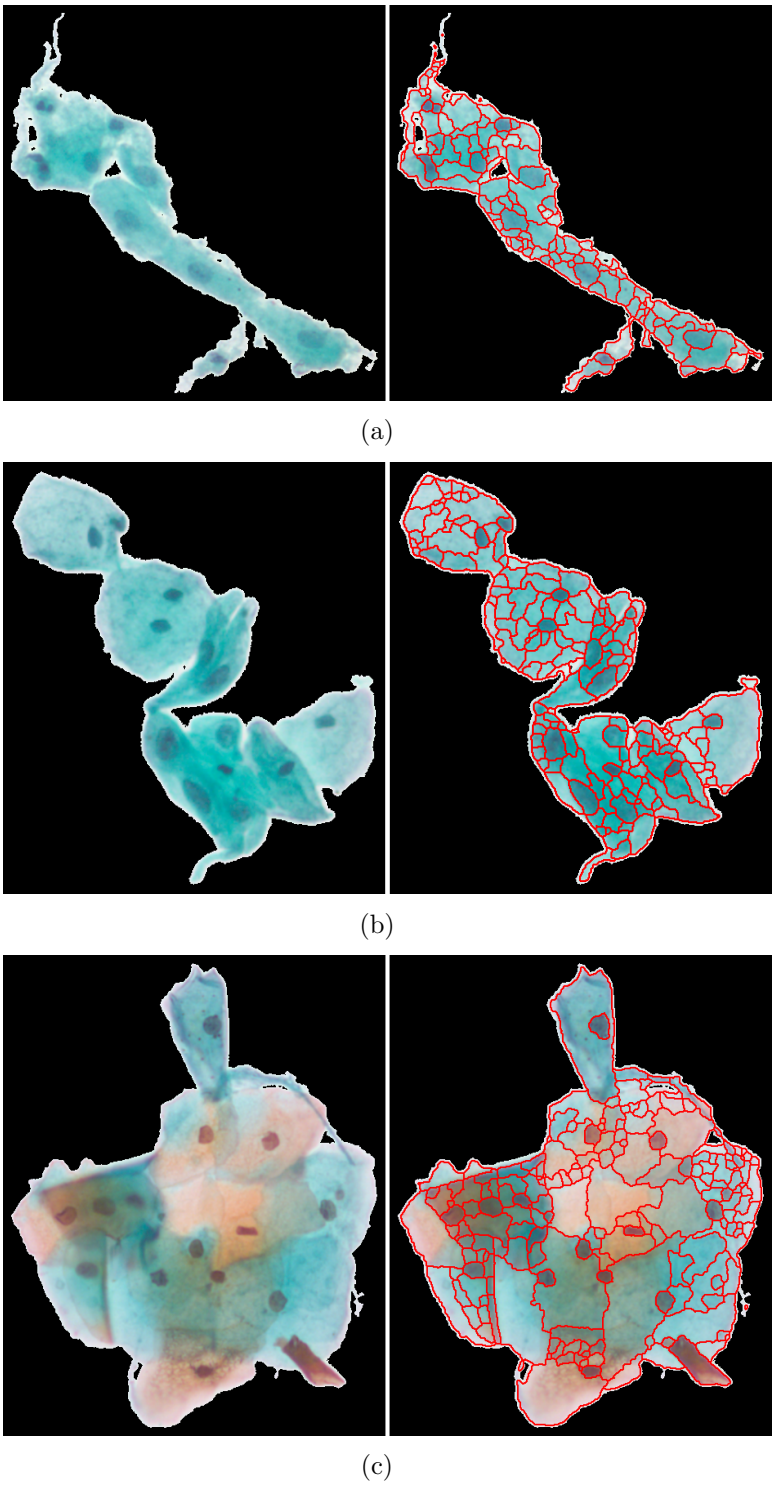


(a)

(b)

(c) (d)

Figure 5.7: Segmentation results for example images from Hacettepe data.

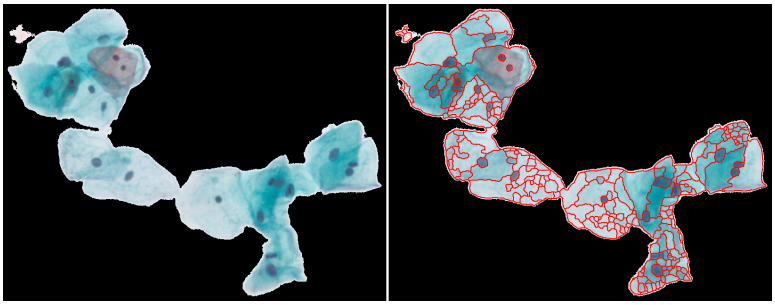


(a)

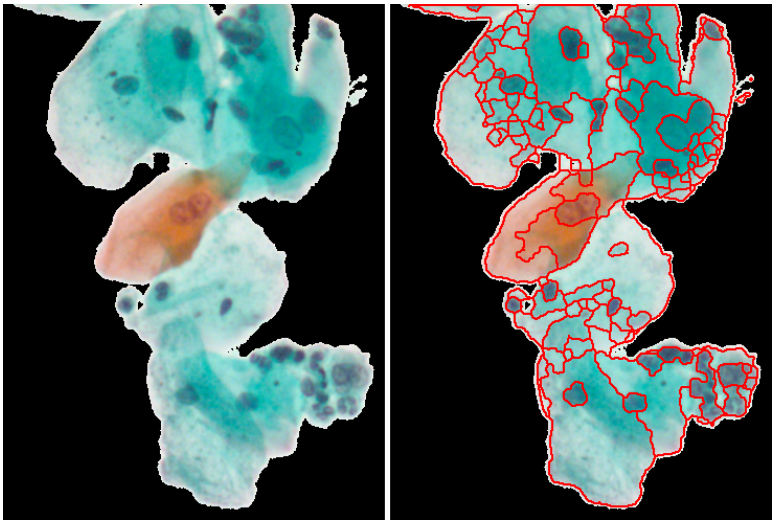
(b)

(c)

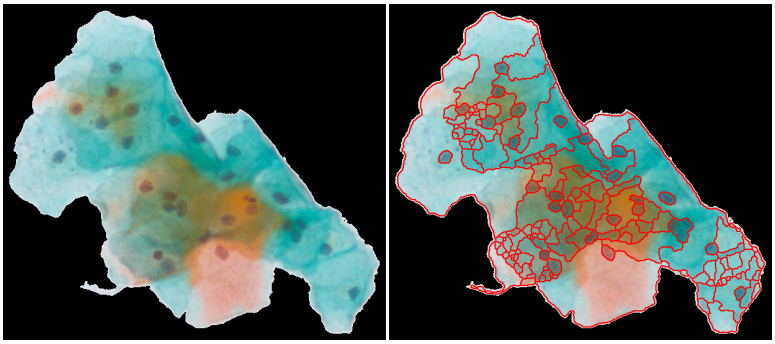
Figure 5.8: Segmentation results for example images from Hacettepe data.



(a)

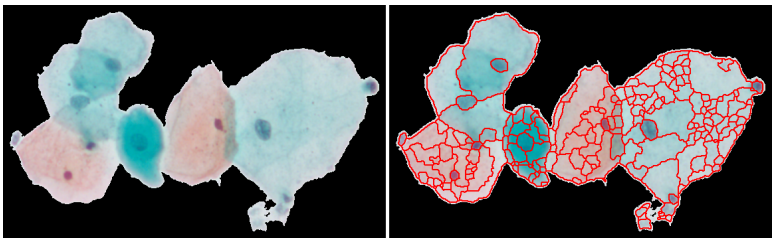


(b)

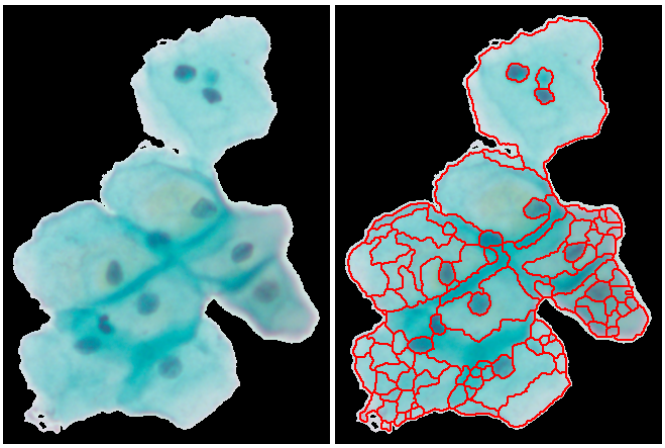


(c)

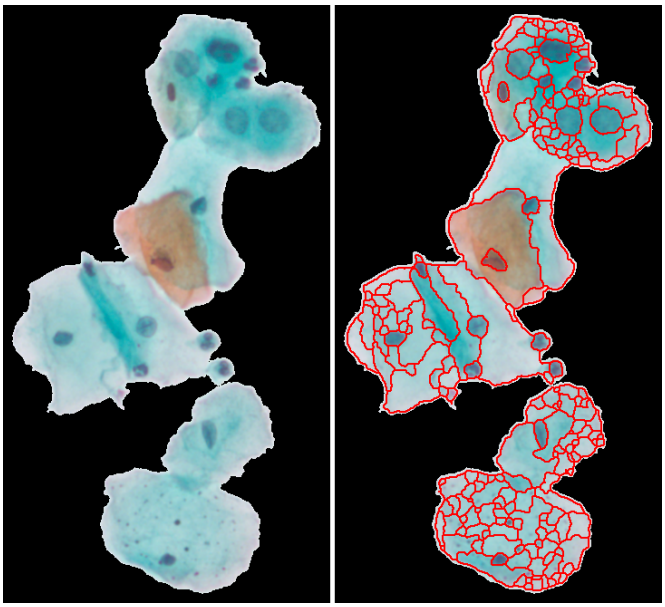
Figure 5.9: Segmentation results for example images from Hacettepe data.



(a)



(b)



(c)

Figure 5.10: Segmentation results for example images from Hacettepe data.

generic and it also allows to employ other heuristics while defining meaningfulness of a region.

Below, we enumerate occasions in which segments of true nucleus regions cannot be obtained using our method.

• True segment of a nucleus region may never appear in a hierarchical tree due to noisy texture of the nucleus region (see the bottom nucleus in Figure 5.11 (a)) or poor contrast on the nucleus contour resulting from inadequate focus of the microscope during image acquisition (see Figure 5.11 (b)).

• True nucleus region appears in a hierarchical tree but its ancestor at a higher level is found to be more meaningful because of the homogeneity factor. For example, the region associated with the nucleus of the bottom cell in the image of Figure 5.11 (c) is first merged with the region associated with the cytoplasm area of that cell. Then, the whole cell is combined with the cytoplasm area of the overlapping cells at the upper part. The homogeneity measure of the whole cell becomes much higher than the homogeneity of its nucleus region such that instead of the nucleus region, the whole cell region is selected from the hierarchy.

• After true nucleus region is formed in a hierarchy, it combines with a small noisy segment near it. Then, this resulting region merge with the cytoplasm area of the cell. The homogeneity factor is much higher for the resulting region compared to the circularity measure of the true nucleus region. Thus, the union of the true nucleus structure and the noisy segment is selected from the hierarchy (see the rightmost nucleus of the image in Figure 5.11 (d)).

Figure 5.12 shows the segmentation result for an example image consisting of many overlapping noisy cells. Most of the nucleus regions are obtained in the resulting segmentation whereas some of them cannot be captured because of the reasons explained above. We propose a generic segmentation method and it allows to add new components to the meaningfulness measure like a component involving size feature. However, there is no obvious solution while using heuristics and it can be elaborated as a future work.

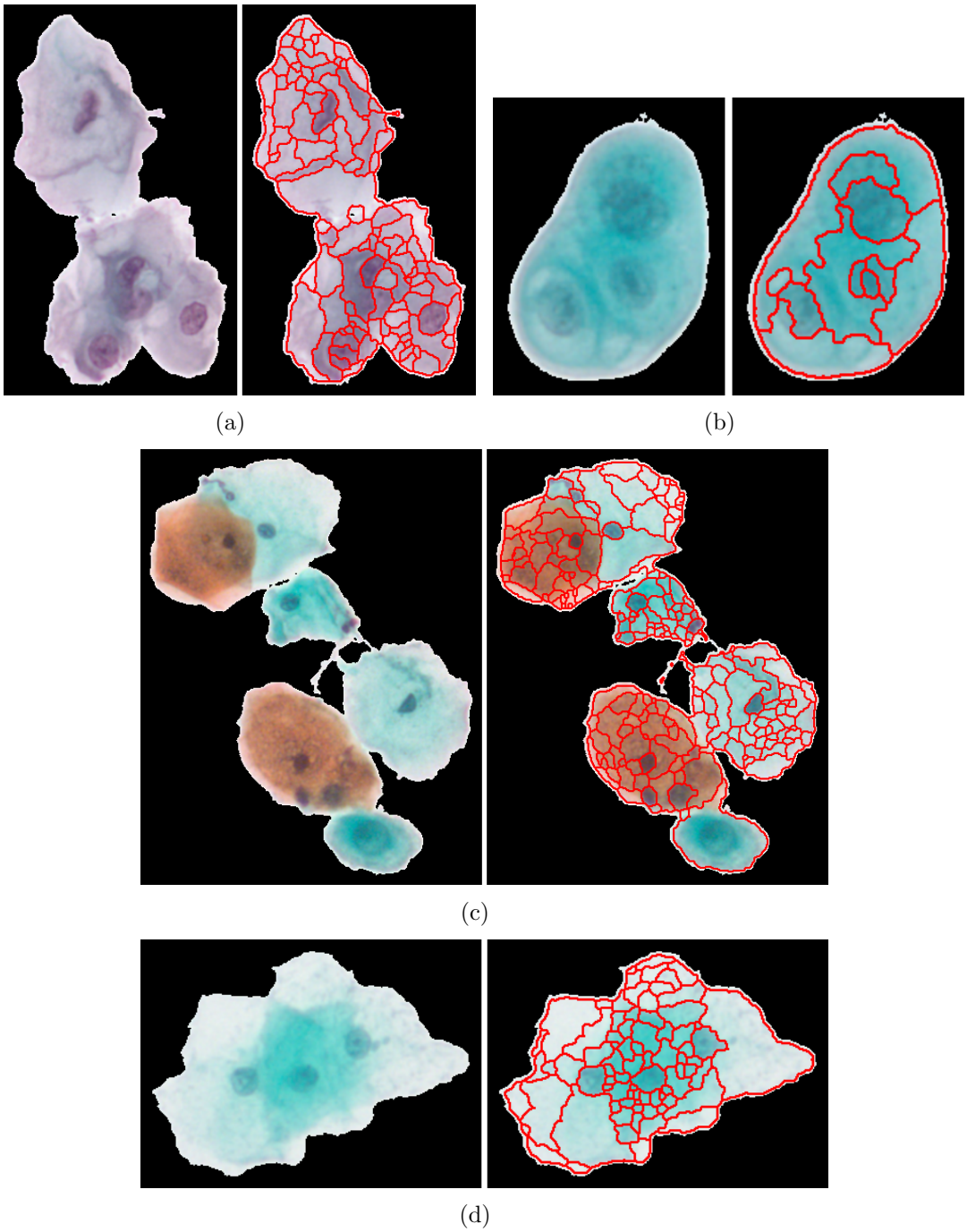
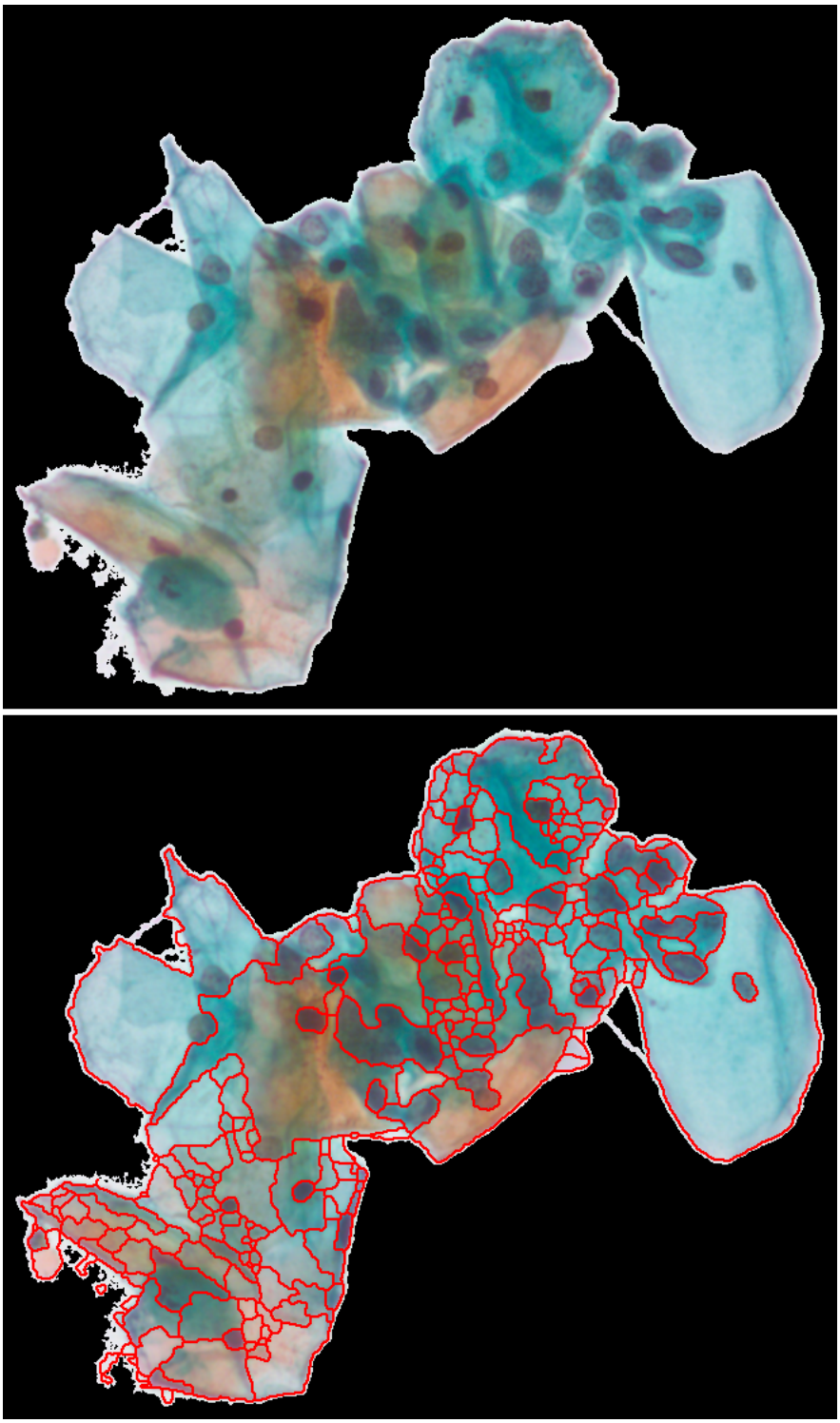


Figure 5.11: Problematic segmentation results for example images from Hacettepe data.



(a)

Figure 5.12: Segmentation result for an example image consisting of many overlapping noisy cells.

As a final matter, a cervical cell rarely has a double-nucleus whereas it frequently arises from two overlapping cells whose nuclei also overlap. Our segmentation method assumes that nucleus regions are circular so that the segmentation may not result in a corresponding region for a double-nucleus due to its shape. Even if we achieve to segment a double-nucleus, it may be later classified as cytoplasm because of its shape and size. Moreover, segmenting and classifying a double-nucleus properly do not lead to a solution at all. If this double-nucleus stems from two overlapping cells, then the number of cells located at the same cell region will be computed one less than the actual number such that the cytoplasm features of the cells become less accurate. Keeping in mind these problems, we will elaborate the case of double-nucleus as a future work.

5.1.3 Nucleus and Cytoplasm Classification

The segments resulting from the segmentation of a cell region are classified as nucleus or cytoplasm area based on their four different features, namely, size, mean intensity, eccentricity, and homogeneity measurements. We first form a data set composed of the corresponding feature vectors of 1452 nucleus regions and 7726 cytoplasm regions which are selected from the segmentation results of the Hacettepe data set. After experimenting with different classifiers on the collected data set, we decided to use the combined classifier based on the sum of posterior probabilities (see Table 3.1). Further details can be found in Section 3.3.

5.2 Ranking of Cervical Cells

In this section, we first describe the statistics of rank-order correlation coefficients and kappa coefficients that we use to evaluate the ranking step of our method. Then, we present the conducted experiments and compare their results based on the given statistics. The experiments are performed on the Herlev data that contain ground truth.

Table 5.2: An example ranking scenario.

Cells	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
Class labels	1	1	1	2	2	2	3	3
Initial ranking	1	2	3	4	5	6	7	8
Ground truth ranking R	2	2	2	5	5	5	7.5	7.5
Our ranking result S	2	2	5	5	2	7.5	5	7.5

5.2.1 Rank-Order Correlation Coefficients

In order to evaluate the ranking step of our algorithm, we use the following methodology. We first select a set containing a total number of N cells from multiple classes in the Herlev data. Either all classes or some of the classes can be used in a single experiment. Then, the selected cells are ranked according to their class labels where the class labels determine their corresponding abnormality degree as previously explained in Section 1.3.1. Note that this ranking is the ideal one that we want to achieve because we aim to order the cells according to their abnormality degree. In this way, we obtain the ground truth ranking R of the cells where we know the rank R_i of each cell $x_i, i = 1, .., N$. An example set of cells and their class labels are given in Table 5.2 where x_1, x_2, x_3 are assumed to belong to class 1, x_4, x_5, x_6 are assumed to belong to class 2 and x_7, x_8 are assumed to belong to class 3. Ranks of the cells with the same class label should be the same so we assign these cells to the mean of their initial ranks and obtain the ground truth ranking R.

Suppose that our algorithm ranks these cells in the order of x_1 x_5 x_2 x_4 x_3 x_7 x_6 x_8. Since we propose to order the cells according to their abnormality degree, we can say that our method labels the first three cells, namely x_1 x_5 x_2, as class 1, the next three cells, namely x_4 x_3 x_7, as class 2, and the last two cells, namely x_6 x_8, as class 3. When we calculate the rankings of the cells based on these class associations, we obtain our ranking result S shown in Table 5.2 where each cell x_i has a corresponding rank $S_i, i = 1, .., N$.

The Spearman rank-order correlation coefficient r_s and the sum of squared

difference of ranks D are used in order to measure the agreement between the ground truth ranking R and our ranking result S [30]. The Spearman rank-order correlation coefficient r_s is defined as

$$r_s = \frac{\sum_{i=1}^{N}(R_i - \overline{R})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{N}(R_i - \overline{R})^2}\sqrt{\sum_{i=1}^{N}(S_i - \overline{S})^2}} \tag{5.1}$$

where $\overline{R}$ and $\overline{S}$ are the mean of R_i's and S_i's, respectively, and the sum of squared difference of ranks D is defined as

$$D = \sum_{i=1}^{N}(R_i - S_i)^2. \tag{5.2}$$

The sign of the Spearman rank-order correlation coefficient indicates the direction of the association between R and S. A Spearman correlation of zero indicates that there is no tendency for S to either increase or decrease when R increases. When R and S are highly correlated, the Spearman correlation increases in magnitude. On the other hand, we expect the sum of squared difference of ranks decrease when the correlation between R and S increases.

5.2.2 Kappa Coefficients

The weighted kappa coefficient κ_w provides a measure of agreement between two raters who classify observations into one of several categories. It is different from simple percent agreement in the sense that it considers the agreement occurring by chance.

Suppose that two raters classify each of N observations into one of g categories. Then, we obtain a confusion matrix n where n_{ij} represents the number of observations that have been classified as belonging to category i by the first rater and the category j by the second rater. We also define a weight matrix w where a weight w_{ij} between 0 and 1 is selected for each n_{ij}. The weight w_{ij} determines the degree of similarity between two categories i and j. Hence, the weights on the diagonal of w are selected as 1 whereas the weights w_{ij} with highly different

categories i and j are given close to or equal to 0. The weighted relative observed agreement among raters is obtained as

$$p_{o(w)} = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{g} w_{ij}\, n_{ij}. \tag{5.3}$$

The weighted relative agreement expected just by chance is estimated by

$$p_{e(w)} = \frac{1}{N^2} \sum_{i=1}^{g} \sum_{j=1}^{g} w_{ij}\, r_i\, c_j \tag{5.4}$$

where $r_i = \sum_{j=1}^{g} n_{ij}$ and $c_j = \sum_{i=1}^{g} n_{ij}$. Then, the weighted kappa coefficient which may be interpreted as the chance-corrected weighted relative agreement is given by

$$\kappa_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}. \tag{5.5}$$

The weighted kappa coefficient becomes Cohen's kappa coefficient κ when the weights w_{ij} are set to 0 for $i \neq j$ which means that all categories are equally different from each other. The kappa coefficients have a maximum of 1 when the agreement between two raters are perfect whereas a value of 0 indicates no agreement better than chance. Negative values show worse than chance agreement.

In order to evaluate the ranking of our method, we use both of the kappa coefficients as follows. Suppose that we rank a set of N cells from g different classes where there are k_i cells from each class $i \in \{1, 2, ..., g\}$. As explained in the previous section, we assume that our ranking method classifies the first k_1 cells of the ordering as belonging to class 1, the next k_2 cells of the ordering as belonging to class 2, and so on. Similarly, the last k_g cells of the ordering are classified as belonging to class g. Since the ground truth labels of the cells are also known, we calculate the corresponding confusion matrix to be used in the

kappa statistics. Here, we determine the weight matrix of κ_w intuitively as

$$\begin{bmatrix} 1 & 0.5 & 0 & 0.25 & 0.25 & 0 & 0 \\ 0.5 & 1 & 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0 & 1 & 0.5 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0 & 0.5 & 1 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.25 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0 & 0.25 & 0.5 & 0.5 & 1 \end{bmatrix}$$

by considering the difference between each pair of classes.

Among the four ranking criteria, higher values of r_s, κ and κ_w and lower values of D indicate better agreement.

5.2.3 Experimental Results

In this part, we describe the experiments conducted for evaluating our ranking of cervical cells. We compare the experimental results based on the given statistics, namely, the Spearman rank-order correlation coefficient r_s, the sum of squared difference of ranks D, the Cohen's kappa coefficient κ and the weighted kappa coefficient κ_w.

We perform the experiments on different settings whose details are given below. For each setting, the experiment is conducted using both all of the features (14 features described in Section 4.1) and the nucleus features (9 features, namely nucleus area, nucleus brightness, nucleus longest diameter, nucleus shortest diameter, nucleus elongation, nucleus roundness, nucleus perimeter, nucleus maxima/minima). The nucleus features are obtained by removing the cytoplasm features from all of the features.

• *Case* 1: We order all of the cells in the whole data using the first criterion of the optimal leaf ordering algorithm which maximizes the sum of similarities between adjacent leaves.

• *Case* 2: We order all of the cells in the whole data using the second criterion

of the optimal leaf ordering algorithm which maximizes the sum of the similarities between every leaf and the leaves in its adjacent clusters. Hereafter, we will use the second criterion for the optimal leaf ordering algorithm because it provides better results using both all of the features and the nucleus features as shown in Table 5.3.

• *Case* 3: We order the cells of all classes except the columnar cells. The columnar cells are rarely encountered in the images of the Hacettepe data and we will not include the columnar cells in the experiments from now on.

• *Case* 4: In this setting, we use the cells of all classes except columnar, severe dysplasia and carcinoma in situ classes. In this way, we aim to evaluate our performance when we are given Pap test images of the patients at the early stage of the disease.

• *Case* 5: In this setting, the cells of all classes except columnar, mild dysplasia and moderate dysplasia classes are ordered.

• *Case* 6: In this setting, we first group the superficial squamous and the intermediate squamous classes into a single class that we call normal. Then, we also group the mild dysplasia,the moderate dysplasia, the severe dysplasia and the carcinoma in situ classes into a single class named abnormal.

• *Case* 7: In this setting, we again group the superficial squamous and the intermediate squamous classes into a single normal class. The mild dysplasia and the moderate dysplasia classes are grouped into a class that we call early-abnormal. We group the severe dysplasia and the carcinoma in situ classes into a class named abnormal.

Table 5.3 summarizes the experimental results obtained for different settings. Using all of the features, we obtain better agreement between the ground truth and the ranking result of our algorithm. The performance of our ranking method improves when there are no columnar cells in the input data. Dropping columnar cells does not lead to an unrealistic situation because we observe that the Pap test images of the Hacettepe data rarely include columnar cells. We obtain almost

Table 5.3: The experimental results for the ranking of cervical cells obtained for different settings.

	Using all features				**Using only nucleus features**			
	r_s	D	κ	κ_w	r_s	D	κ	κ_w
Case 1	0.675	40474731.5	0.265	0.328	0.387	76462544.0	0.066	0.149
Case 2	0.704	36853826.0	0.282	0.338	0.388	76410864.0	0.068	0.155
Case 3	0.845	13573249.5	0.431	0.559	0.211	69464205.5	0.055	0.140
Case 4	0.785	3402208.0	0.509	0.581	0.646	5613168.0	0.348	0.484
Case 5	0.709	5166073.0	0.382	0.604	0.656	6104719.0	0.385	0.536
Case 6	0.848	6036849.0	0.848	0.848	0.806	7713751.5	0.806	0.806
Case 7	0.814	14534743.5	0.716	0.764	0.254	58304191.0	0.058	0.253

perfect agreement by grouping the data into normal and abnormal classes for which both kappa coefficients κ and κ_w are calculated as greater than 0.8. This supports the conjecture that the cervical cells can be grouped according to their abnormality degree using our ranking method. Moreover, we achieve the substantial agreement when we group the mild dysplasia and the moderate dysplasia as well as the severe dysplasia and the carcinoma in situ classes. This setting is appropriate because the severe dysplasia and the carcinoma in situ classes are considered as very similar [26].

In Figures 5.13 to 5.15, three different orderings for a set of cells are given where we manually observe that the first order is superior to the second one and the second order is better than the third one. Now, we aim to verify this relationship using the statistical coefficients that we use to measure the agreement between our ranking and the ground truth. The corresponding statistics of the orderings are given in Figures 5.13 to 5.15 and they are consistent with our observation. For example, all of the coefficients indicates the best agreement for the first ordering and the worst agreement for the last ordering. Moreover, the weighted kappa coefficient is less than 0 for the last ordering meaning that the agreement is worse than chance which is consistent with the fact that this ordering is generated randomly.

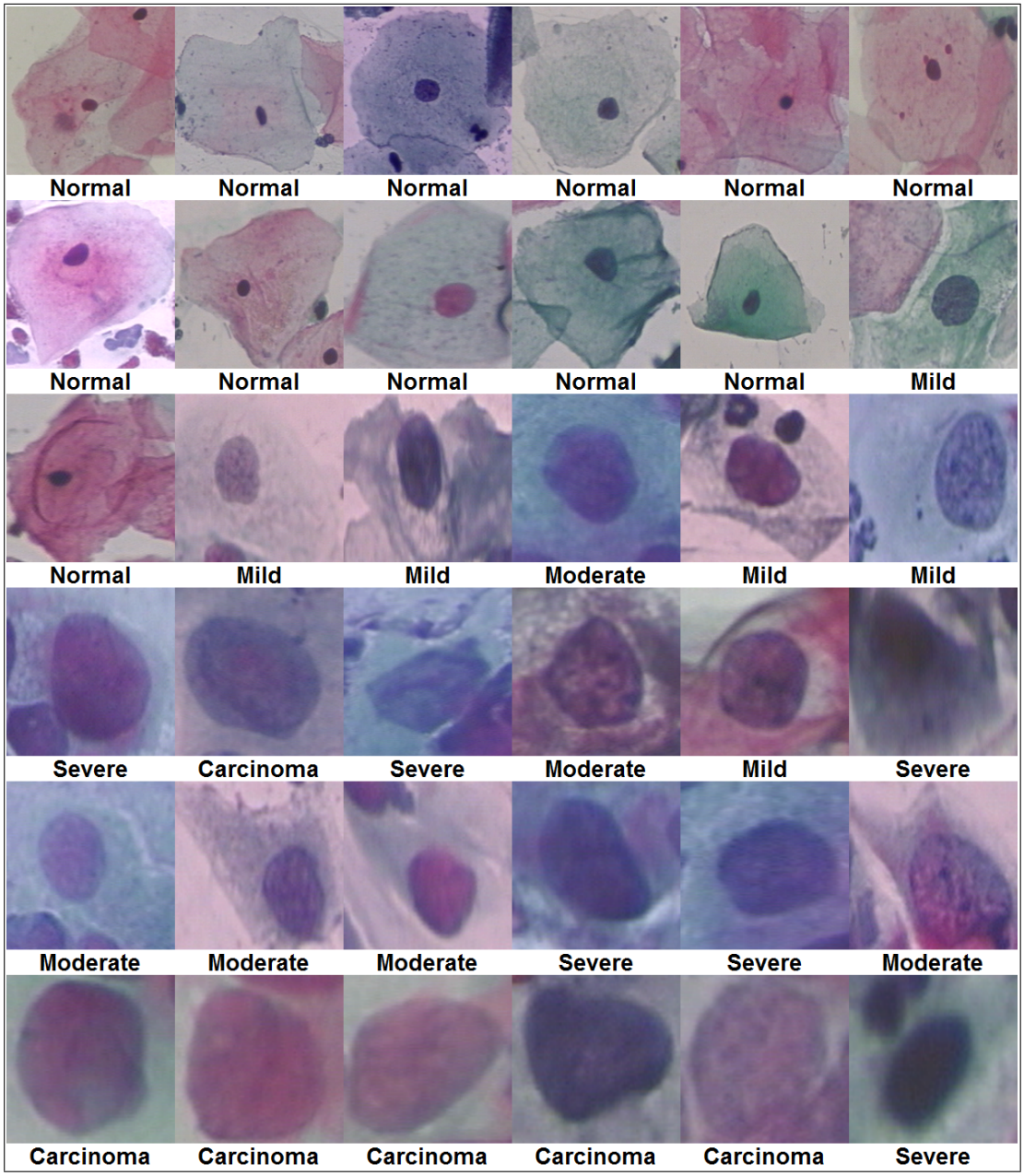


Figure 5.13: The first ordering of the cells where $r_s = 0.895$, $D = 792$, $\kappa = 0.466$ and $\kappa_w = 0.614$. (The cell images are resized to have the same width and height so their relative size is not proper.)

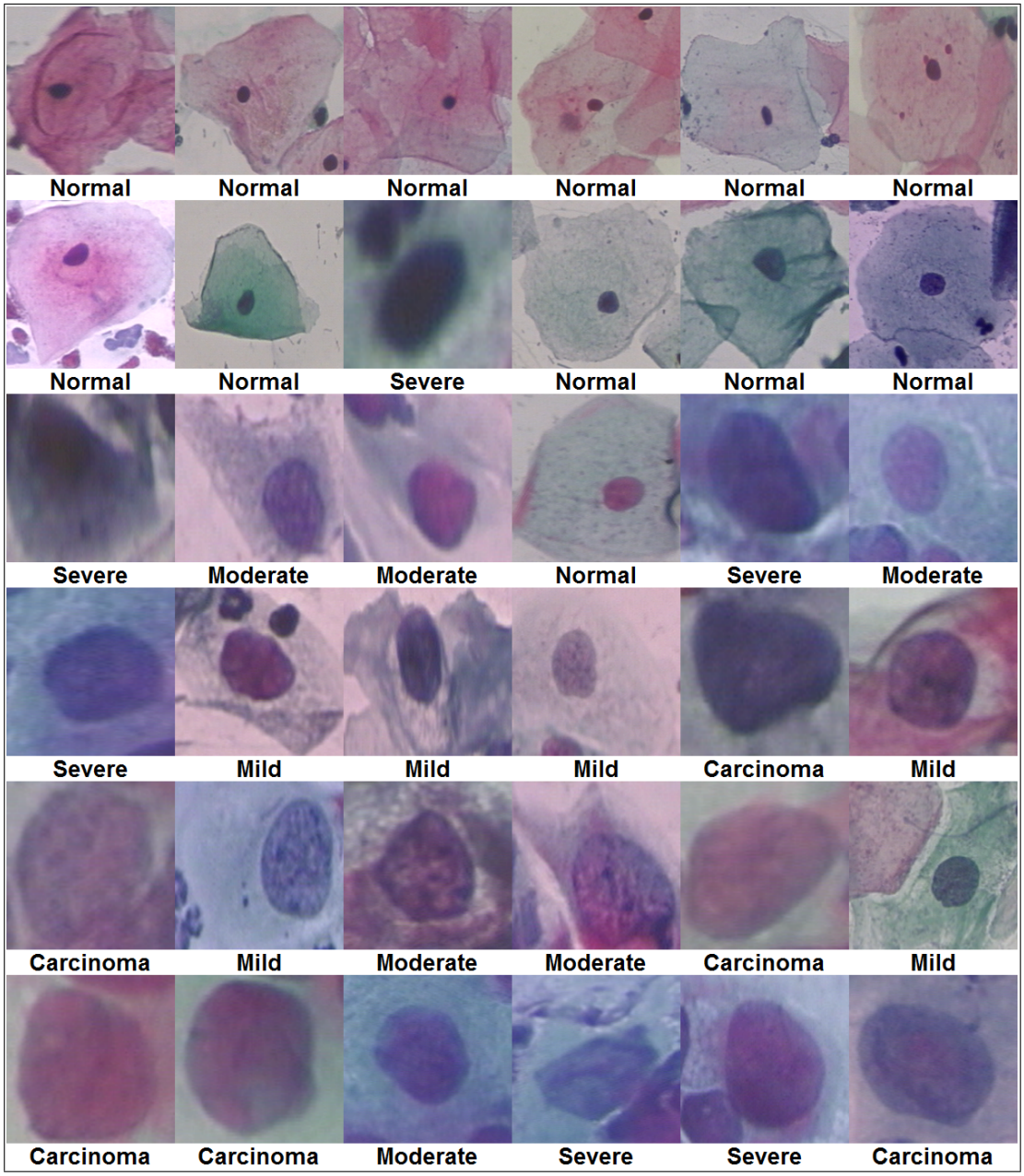


Figure 5.14: The second ordering of the cells where $r_s = 0.771$, $D = 1728$, $\kappa = 0.266$ and $\kappa_w = 0.417$. (The cell images are resized to have the same width and height so their relative size is not proper.)

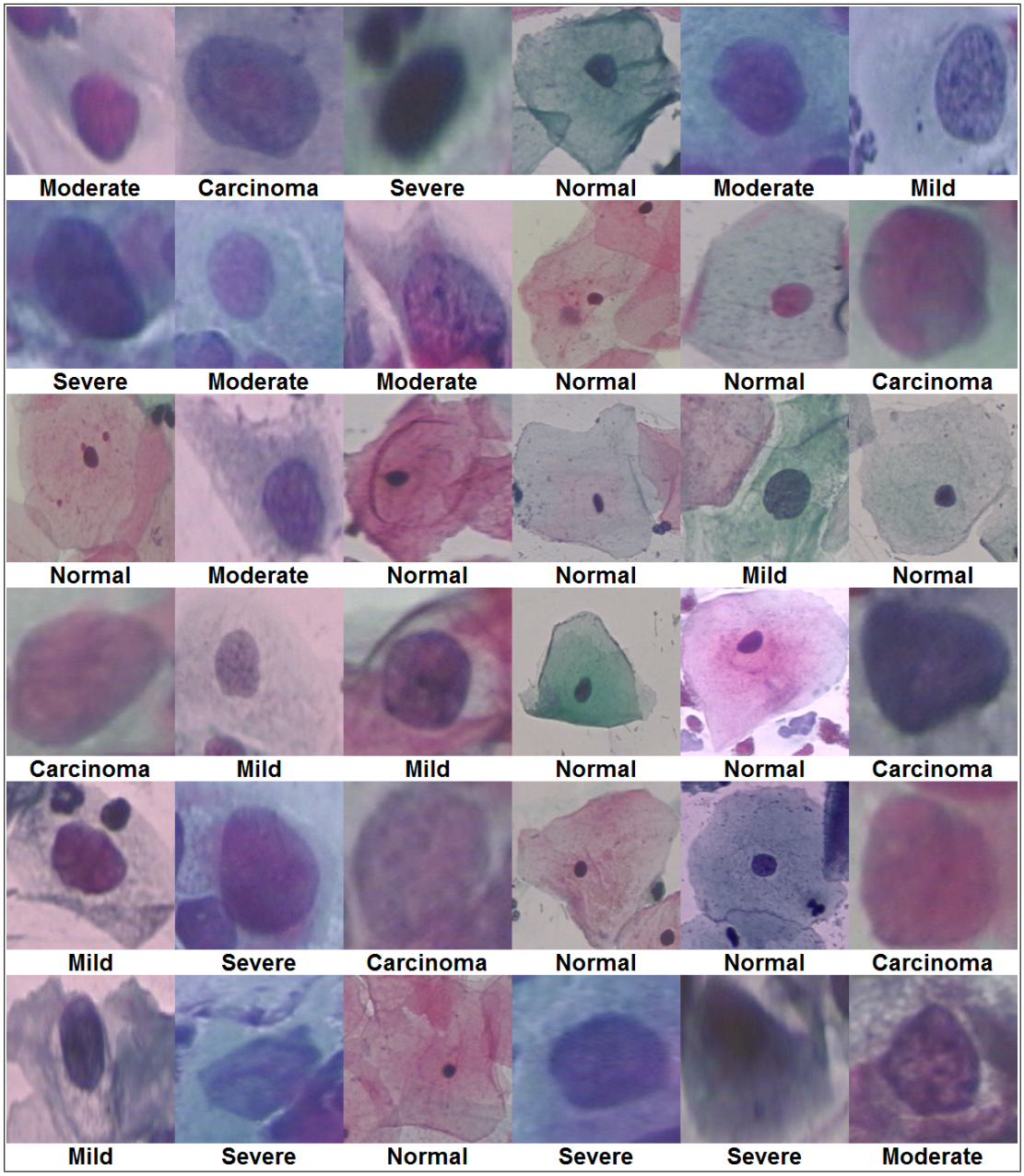


Figure 5.15: The third ordering of the cells where $r_s = 0.038$, $D = 7272$, $\kappa = -0.100$ and $\kappa_w = -0.054$. (The cell images are resized to have the same width and height so their relative size is not proper.)

Chapter 6

Conclusion and Future Work

In this thesis, we presented a computer-assisted screening system which aimed to help cyto-technicians by sorting cells in a Pap slide according to their abnormality degree. In this system, cells along with their nuclei were first located using a segmentation procedure on an image taken using a microscope. Then, we extracted features describing these segmented cells. Lastly, the cells were ordered according to their abnormality degree based on the extracted features.

Different from the related studies involving images of a single cell, our three-phase segmentation procedure could handle images containing overlapping cells. Thresholding method was used as the first phase to extract background regions for obtaining remaining cell regions. In the second phase, we segmented the cell regions by a method following the general framework of the segmentation approach developed by Akçay and Aksoy [1]. The main difference and advantage of our method stems from being a non-parametric hierarchical segmentation algorithm that uses the spectral and shape information as well as the gradient information. The last phase aimed to partition the cell region into true structures of each nucleus and the whole cytoplasm area by classifying the final segments as nucleus or cytoplasm region. Our experiments showed that the proposed segmentation procedure performed well for images of overlapping cells as well as a single cell.

In order to rank cells, we first performed hierarchical clustering on 14 different

cell features. The initial ordering of the cells was determined as the leaf ordering of the constructed hierarchical tree because hierarchical clustering groups the closest pair of clusters at each step and the most related cells or cell groups become adjacent in the linear ordering of the tree. Then, this initial ordering was improved by applying the optimal leaf ordering algorithm [6]. Our experiments showed that the cervical cells could be grouped according to their abnormality degree using our ranking method.

A cervical cell rarely has a double-nucleus whereas it frequently arises from two overlapping cells whose nuclei also overlap. Our segmentation procedure may not generate a corresponding segment for a double-nucleus due to its shape. When this double-nucleus stems from two overlapping cells, the features of the cells located at the same cell region become less accurate. As a future work, we will elaborate on the case of cells with double-nucleus. Another future work is the registration of low and high magnification images in order to enhance the cell features such as texture of nucleus region.

Bibliography

[1] H. Akcay and S. Aksoy. Automatic detection of geospatial objects using multiple hierarchical segmentations. *IEEE Transactions on Geoscience and Remote Sensing*, 46(7):2097–2111, 2008.

[2] A. Alizadeh, M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.

[3] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745, 1999.

[4] P. Bamford and B. Lovell. A water immersion algorithm for cytological image segmentation. In *Proceedings of the APRS Image Segmentation Workshop*, pages 75–79. Citeseer.

[5] P. Bamford and B. Lovell. Unsupervised cell nucleus segmentation with active contours. *Signal Processing*, 71(2):203–213, 1998.

[6] Z. Bar-Joseph, D. Gifford, and T. Jaakkola. Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics-Oxford*, 17:22–29, 2001.

[7] M. Bazoon, D. Stacey, C. Cui, and G. Harauz. A hierarchical artificial neural network system for the classification of cervical cells. In *1994 International Conference on Neural Networks*, 1994.

[8] H. Causton, B. Ren, S. Koh, C. Harbison, E. Kanin, E. Jennings, T. Lee, H. True, E. Lander, and R. Young. Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of the Cell*, 12(2):323–337, 2001.

[9] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/c̃jlin/libsvm.

[10] L. Cohen. On active contour models and balloons. *Computer Vision Graphics and Image Processing: Image Understanding*, 53(2):211–218, 1991.

[11] I. Dagher and K. Tom. WaterBalloons: A hybrid watershed Balloon Snake segmentation. *Image and Vision computing*, 26(7):905–912, 2008.

[12] M. Desai. Role of automation in cervical cytology. *Diagnostic Histopathology*, 15(7):323–329, 2009.

[13] E. Dougherty and R. Lotufo. *Hands-on Morphological Image Processing*. SPIE press, 2003.

[14] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley New York, 2001.

[15] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

[16] P. Huang and Y. Lai. Effective segmentation and classification for HCC biopsy images. *Pattern Recognition*, 2009.

[17] N. C. Institute. Health Report Fiscal Years 2005-2006, 2007.

[18] J. Jahne and H. Horst. *Computer vision and applications, a guide for students and practitioners*. Academic Press, 2000.

[19] A. Kale, S. Aksoy, and S. Onder. Pap Smear Test Goruntulerinde Hucre Cekirdeklerinin Bolutlenmesi (in Turkish). In *17. IEEE Sinyal Isleme ve Iletisim Uygulamalari Kurultayi*, 2009.

[20] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contour models. *International Journal of Computer Vision*, 1(4):321–331, 1988.

[21] S. Keerthi and C. Lin. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*, 15(7):1667–1689, 2003.

[22] J. Kittler and J. Illingworth. Minimum error thresholding. *Pattern Recognition*, 19(1):41–47, 1986.

[23] E. Martin. Pap-smear classification. *Master's Thesis, Technical University of Denmark: Oersted-DTU, Automation*, 2003.

[24] V. Meas-Yedid, S. Tilie, and J.-C. Olivo-Marin. Color image segmentation based on markov random field clustering for histological image analysis. In *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02) Volume 1*, page 10796, Washington, DC, USA, 2002. IEEE Computer Society.

[25] F. Meyer. Topographic distance and watershed lines. *Signal Processing*, 38(1):113–125, 1994.

[26] J. Norup. Classification of pap-smear data by transductive neuro-fuzzy methods. *Master's Thesis, Technical University of Denmark: Oersted-DTU, Automation*, 2005.

[27] L. O'gorman and A. Sanderson. The converging squares algorithm: An efficient method for locating peaks in multidimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(3):280–288, 1984.

[28] G. Paschos. Perceptually uniform color spaces for color texture analysis: an empirical evaluation. *IEEE Transactions on Image Processing*, 10(6):932–937, 2001.

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

[30] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. Numerical recipes in C: the art of scientific programming. *Cambridge U. Press, Cambridge, England*, 1992.

[31] J. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27:221–234.

[32] S. Shah. Automatic cell image segmentation using a shape-classification model. *Machine Vision and its Applications*, pages 428–432, 2007.

[33] P. Soille. *Morphological Image Analysis: Principles and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003.

[34] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297, 1998.

[35] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander, and T. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6):2907, 1999.

[36] N. Theera-Umpon. White blood cell segmentation and classification in microscopic bone marrow images. *Lecture Notes in Computer Science*, 3614:787, 2005.

[37] L. Vincent and P. Soille. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):583–598, 1991.

[38] R. Walker, P. Jackway, B. Lovell, and I. Longstaff. Classification of cervical cell nuclei using morphological segmentation and textural feature extraction. In *Proceedings of the 2nd Australian and New Zealand Conference on Intelligent Information Systems*, pages 297–301, 1994.

[39] E. Weisstein. Ellipse. *From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/Ellipse.html.*

[40] H. Wu, J. Barba, and J. Gil. A parametric fitting algorithm for segmentation of cell images. *IEEE Transactions on Biomedical Engineering*, 45(3):400–407, 1998.

[41] H. Wu, J. Barba, and J. Gil. Iterative thresholding for segmentation of cells from noisy images. *Journal of Microscopy*, 197(3):296, 2000.

[42] H. Wu, J. Gil, and J. Barba. Optimal segmentation of cell images. *IEE Proceedings-Vision, Image, and Signal Processing*, 145(1):50–56, 1998.

[43] C. Xu and J. Prince. Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Irocessing*, 7(3):359–369, 1998.

[44] S. Yang-Mao, Y. Chan, and Y. Chu. Edge Enhancement Nucleus and Cytoplast Contour Detector of Cervical Smear Images. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 38(2):353–366, 2008.

[45] A. Zijdenbos, B. Dawant, R. Margolin, and A. Palmer. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Transactions on Medical Imaging*, 13(4):716–724, 1994.