

# UNSUPERVISED SEGMENTATION AND ORDERING OF CERVICAL CELLS

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING  
AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE  
OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
MASTER OF SCIENCE

By  
Nermin Samet  
July, 2014

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.

---

Assoc. 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.

---

Assoc. Prof. Dr. ıđdem Gündüz Demir

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. Nazlı İıkizler-Cinbiş

Approved for the Graduate School of Engineering and Science:

---

Prof. Dr. Levent Onural  
Director of the Graduate School

# ABSTRACT

## UNSUPERVISED SEGMENTATION AND ORDERING OF CERVICAL CELLS

Nermin Samet

M.S. in Computer Engineering

Supervisor: Assoc. Prof. Dr. Selim Aksoy

July, 2014

Cervical cancer is the second most common cause of cancer death among women worldwide, and it can be prevented if it is detected and treated in the pre-cancerous stages. Pap smear test is a common, efficient and easy manual screening examination technique which is used to detect dysplastic changes in cervical cells. However, manual analyses of thousands of cells in Pap smear test slides by cyto-technicians is difficult, time consuming and subjective. To overcome these problems, we aim to automate the screening process and provide an ordered nuclei list to help the cyto-experts. Automating the screening procedure has been a longstanding challenge because of complex cell structures where current methods in the literature mostly consider the problem as the segmentation of single isolated cells and leave real challenges of Pap smear images such as poor contrast, inconsistent staining, and unknown number of cells unaddressed.

We propose an unsupervised method to accurately segment the nuclei and order them according to their abnormality degree in Pap smear images. The method first uses a multi-scale hierarchical segmentation algorithm for accurate identification of the nuclei. The Pap smear images captured at high level magnification have more detailed texture but worse contrast. Contrast is an important property for segmentation and detailed texture is an important property for feature extraction. Therefore, as a solution to the segmentation problem, we proceed in two steps. First, we segment the Pap smear images at low (20x) magnification and eliminate non-nucleus regions based on several features. Then, we switch to high (40x) magnification and obtain a more detailed segmentation of the remaining nuclei. Following segmentation, we extract features for each resulting nucleus. Unlike related works that require a learning phase for classification, our method performs an unsupervised ordering of the nuclei based on features extracted at 40x magnification. We compare different ordering algorithms for ranking the

nucleus regions according to their abnormality degrees.

We evaluate our segmentation and ordering methods using two data sets. Our results show that the proposed method provides promising results for both segmentation and ordering steps.

*Keywords:* Pap smear test, Pap smear image analysis, Cervical cell segmentation, Multi-scale segmentation, Ordering, Cell grading.

## ÖZET

# SERVİKS HÜCRELERİNİN ÖĞRETİCİSİZ OLARAK BÖLÜTLENMESİ VE SIRALANMASI

Nermin Samet

Bilgisayar Mühendisliği, Yüksek Lisans

Tez Yöneticisi: Doç. Dr. Selim Aksoy

Temmuz, 2014

Serviks kanseri dünya üzerinde kadınlarda en sık görülen ve kanser ölümlerine sebep olan ikinci kanser çeşididir. Serviks kanseri prekanseröz aşamalarda erken teşhis ve tedavi ile önlenabilmektedir. Pap smear testi, serviks hücrelerinde meydana gelen displastik değişiklikleri belirlemek üzere kullanılan yaygın, etkili ve kullanımı kolay manuel bir tarama yöntemidir. Ancak Pap smear testlerinde bulunan binlerce hücrenin sitologlar tarafından manuel olarak analiz edilmesi zorlu, zaman alan ve gözlemci özneliği içeren bir süreçtir. Çalışmamızda, bu sorunların üstesinden gelmek için tarama işlemini otomatikleştirmeyi ve sitologlara yardımcı olacak hücrelerin sıralanmış listesini sağlamayı amaçladık. Tarama sürecini otomatikleştirme, karmaşık hücre yapılarından dolayı uzun süreli ve zorlu bir görev olarak durmaktadır. Literatürdeki mevcut yöntemler çoğunlukla problemi tekli ve ayrılmış hücre bölütlemesi olarak ele almakta ve Pap smear test görüntülerinin, zayıf kontrast, tutarsız boyama ve bilinmeyen hücre sayısı gibi gerçek sorunlarına değinmemektedirler.

Bu tezde, Pap smear görüntülerindeki hücrelerin doğru bir biçimde bölütlenmesi ve anormallik derecelerine göre sıralanması için öğreticisiz bir yöntem önerilmektedir. Önerilen yöntem ilk olarak çekirdeklerin doğru bir şekilde elde edilmesi için çoklu-ölçekli hiyerarşik bölütleme algoritması kullanmaktadır. Yüksek büyütme değeri ile çekilen Pap smear görüntüleri daha detaylı doku bilgisine ancak daha kötü kontrast değerine sahiptirler. Kontrast bölütleme aşaması için önemli bir özellik iken, detaylı doku bilgisi öznelik çıkarma aşaması için önemli bir özelliktir. Bu nedenle, çalışmamızda bölütleme problemine bir çözüm olarak, iki aşamada ilerledik. İlk olarak, Pap smear görüntüleri düşük büyütme (20x) seviyesinde bölütlendi ve çıkarılan çeşitli özneliklere dayanarak çekirdek olmayan bölütlenmiş alanlar elendi. Daha sonra, yüksek seviyede (40x) çekilen

Pap smear görüntülerine geçilerek kalan çekirdeklerin daha detaylı bölütlenmesi gerçekleştirildi. Bölütleme aşamasının ardından, elde edilen her çekirdek için öznitelikler çıkarıldı. Literatürdeki sınıflandırma için öğrenme aşaması gerektiren ilgili çalışmalardan farklı olarak, yöntemimiz 40x büyütme oranındaki görüntülerden çıkarılan özniteliklere dayanarak çekirdeklerin öğreticisiz olarak sıralamasını gerçekleştirmektedir. Farklı sıralama algoritmaları, elde edilen çekirdeklerin anormallik derecelerine göre sıralanması üzerinden karşılaştırıldı.

Bölütleme ve sıralama yöntemlerimizi iki veri kümesi kullanarak değerlendirdik. Sonuçlarımız önerilen yöntemlerin hücrelerin hem bölütlenmesi hem de sıralanması aşamasında gelecek vaat eden sonuçlar verdiğini gösterdi.

*Anahtar sözcükler:* Pap smear testi, Pap smear görüntü analizi, Serviks hücre bölütlenmesi, Çoklu-ölçekli bölütleme, Sıralama, Hücre derecelendirmesi.

## Acknowledgement

I would like to thank my advisor Assoc. Prof. Dr. Selim Aksoy for his supervision through my research. I would like to thank to the members of my thesis committee Assoc. Prof. Dr. Çiğdem Gündüz Demir and Assist. Prof. Dr. Nazlı İkizler-Cinbiş for accepting to review my thesis and to be in my thesis committee. I thank to Dr. Sevgen Önder for his consultancy on medical knowledge and providing us the Hacettepe dataset.

I would like to express my gratitude to Dr. Wolfgang Stürzl for providing me summer internship. It has been a great pleasure to work with him and get benefit from his vision and knowledge during my internship at DLR.

I would like to thank Fadime, for always being together. We walk the line together in the good and bad days of the last seven years, best friend ever!

My special thanks go to my dear friend Rabia. She always has been a great friend ever since we began to share a dormitory room when we were 13 years old.

I sincerely thank to all my friends from the RETINA group especially to Caner, Gökhan, Hande, Yiğit, İlker, Acar, Anıl, Sermetcan, Eren and Ahmet. Our enjoyable moments, especially, SIU days and Quick China meetings are always be memorable.

I also would like to thank my amazing friends İbrahim, Tuğba, Neslihan, Melis, Gökhan, Selcen, Oltan, Harun, Ayşegül, Gülcan and Kevser. Oldies but goodies!

I thank to my friends Gülce, Tuğba and Betül, for the amusing time we shared together in the Hacettepe University.

Last but not the least; I would like to thank my beloved family for always believing in me and supporting me spiritually throughout my life. Without them none of them would be possible.

# Contents

- 1 Introduction** **1**
  - 1.1 Problem Definition and Motivation . . . . . 2
  - 1.2 Data Set . . . . . 6
    - 1.2.1 Hacettepe Data Set . . . . . 6
    - 1.2.2 Herlev Data Set . . . . . 7
  - 1.3 Contributions . . . . . 7
  
- 2 Related Work** **12**
  
- 3 Segmentation** **17**
  - 3.1 Segmentation Method . . . . . 18
  - 3.2 Segmentation at 20x Magnification . . . . . 22
  - 3.3 Segmentation at 40x Magnification . . . . . 24
  
- 4 Feature Extraction and Distance Measures** **32**
  - 4.1 Feature Extraction . . . . . 32

<i>CONTENTS</i>	ix
4.2 Distance Measures . . . . .	34
<b>5 Ordering</b>	<b>37</b>
<b>6 Experiments</b>	<b>41</b>
6.1 Evaluation of Segmentation . . . . .	41
6.2 Evaluation of Ordering . . . . .	45
6.2.1 Ordering Results of the Herlev Data Set . . . . .	45
6.2.2 Ordering Results of the Hacettepe Data Set . . . . .	56
6.3 Implementation Settings and Computational Complexity . . . . .	56
<b>7 Conclusion</b>	<b>67</b>

# List of Figures

1.1	An example cell from a Pap smear slide with its background, cytoplasm and nucleus after the staining procedure. . . . .	3
1.2	An example 20x magnification Pap smear image with inconsistent staining, poor contrast, grouped, occluded and overlapped cells. The red circles depict inflammations and other microorganisms. . . . .	4
1.3	Main steps of the proposed automatic segmentation and ordering procedure for the cells in Pap smear Images. . . . .	6
1.4	Three Pap smear images of the same area with the size of $512 \times 512$ , $1024 \times 1024$ and $2048 \times 2048$ respectively correspond to 10x, 20x and 40x magnification levels. . . . .	11
3.1	20x (1st row) and 40x (2nd row) magnification Pap smear images with inconsistent staining, poor contrast and overlapping cells. . . . .	19
3.2	The original image at 20x magnification (A), close up view of the red rectangle at 20x magnification which has more contrast (B) and the corresponding image at 40x magnification which has more detail (C). . . . .	20
3.3	Over segmented 20x magnification Pap smear image result when the segmentation algorithm is applied directly. . . . .	27

3.4	Segmentation steps for a 20x magnification Pap smear image. Raw image (1 <sup>st</sup> row), the same Pap smear image after the segmentation algorithm is applied (2 <sup>nd</sup> row), potential nucleus regions after eliminating the rest of the regions (3 <sup>rd</sup> row). . . . .	28
3.5	An example area from 20x magnification segmented image; (a) before region elimination, (b) after region elimination. . . . .	29
3.6	The overall process of obtaining a segmented nucleus region from a 40x magnification Pap smear image. . . . .	30
3.7	Initial segmentation result of the given 40x magnification nucleus image (a), calculated 40x magnification coarse boundary from 20x magnification nucleus template (b), the merged regions whose 75% overlap with the coarse boundary of the nucleus (c). . . . .	30
3.8	Final segmentation result of 40x magnification Pap Smear Image.	31
4.1	An overview of combining distance matrixes obtained from features	35
6.1	Segmentation results of three Pap smear images at 20x magnification. 1 <sup>st</sup> column shows initial segmented results and 2 <sup>nd</sup> row shows the selected nucleus regions. . . . .	57
6.2	Final segmentation results at 40x magnification. The given images are corresponding pairwise images of 20x magnification Pap smear images shown in Figure 6.1. . . . .	58
6.3	Three Pap smear images that correspond to the same Pap smear slide area at three different focus settings. . . . .	59
6.4	Best performance of the ordering algorithm HC: $k = 0.288$ , $k_w = 0.411$ . The images are resized to the same width and height so the relative sizes of the cells are not proper. . . . .	60

6.5 Best performance of the ordering algorithm OLO:  $k = 0.296$ ,  $k_w = 0.425$ . The images are resized to the same width and height so the relative sizes of the cells are not proper. . . . . 61

6.6 Best performance of the ordering algorithm GW:  $k = 0.328$ ,  $k_w = 0.425$ . The images are resized to the same width and height so the relative sizes of the cells are not proper. . . . . 62

6.7 Best performance of the ordering algorithm TSP:  $k = 0.288$ ,  $k_w = 0.414$ . The images are resized to the same width and height so the relative sizes of the cells are not proper. . . . . 63

6.8 Best performance of the ordering algorithm Chen:  $k = 0.256$ ,  $k_w = 0.386$ . The images are resized to the same width and height so the relative sizes of the cells are not proper. . . . . 64

6.9 Best performance of the ordering algorithm ARSA:  $k = 0.264$ ,  $k_w = 0.395$ . The images are resized to the same width and height so the relative sizes of the cells are not proper. . . . . 65

6.10 Ordering result for the segmented nucleus regions of Figure 6.2(a). The tones of colors represent the similarities between the nucleus regions. . . . . 66

# List of Tables

1.1	Normal Cells . . . . .	9
1.2	Abnormal Cells . . . . .	10
6.1	The ZSI results of three Pap smear images for the ground truth compared to our segmentation. . . . .	42
6.2	Features for the Herlev Data Set . . . . .	49
6.3	HC Ordering Performance . . . . .	50
6.4	OLO Ordering Performance . . . . .	51
6.5	GW Ordering Performance . . . . .	52
6.6	TSP Ordering Performance . . . . .	53
6.7	Chen Ordering Performance . . . . .	54
6.8	ARSA Ordering Performance . . . . .	55
6.9	Best Performance Analyses of the Ordering Algorithms . . . . .	55

# Chapter 1

## Introduction

Cervical cancer is the second leading cause of cancer mortality among women. According to World Health Organization (WHO), every year there are around 530.000 new cases worldwide, and 275.000 of them ends up with death [1]. Cervical cancer usually develops over a long period of time. In this long period which takes years, some early changes occur in the cervix cells. These precancerous changes in cervical cells are known as dysplasia and these dysplastic changes in precancerous cells potentially could develop into cancer. Unfortunately, cervical cancer is mostly unresponsive to treatments at the late stages. However, it is preventable by the treatment of precancerous lesions when the early dysplastic changes occur in the cervix cells [1].

At this point screening plays an important role in detecting these precancerous cells. Among many screening test, the most common screening procedure is Pap smear also known as the Pap smear test which is introduced by Papanicolaou in 1940 [2]. The Pap smear is a test which is used to detect the changes in the cervix cells that are cancer or potentially lead to cancer. This technique aims to detect precancerous and cancerous cells by analyzing colored and stained Pap smear slides. In order to detect abnormal changes in the cervix cells, cytotechnicians analyse these Pap smear slides in laboratories using a microscope under the supervision of a pathologist. They basically examine the cells according to their shape, color, size, nucleus proportion to cytoplasm and finally categorize the cells

according to their abnormality degree.

Since cervical cancer mortality rates have decreased over the past decades with the widespread use of the Pap smear, it is a preferable technique as an effective, economical and simple method [3]. However, manual-screening procedure is open to inaccurate diagnoses and human driven errors. Automating this manual screening procedure could be a plausible solution to avoid these issues. However, automating this procedure is a challenging problem because of the complexities in cervix cell structures. Although a large number of studies have been done on the automatization of the Pap smear test procedure, it is still a manual-screening procedure. In this work, we present an automatic computer-assisted system which segments and orders nuclei in the cells of a Pap smear slide image according to their dysplasia degree in an unsupervised way. With the help of our system cytotechnicians could skip normal cells and focus on the cells with dangerous nuclei. Our procedure consists of two main steps. The first and the most crucial step is the accurate segmentation of nucleus regions and the second step is ordering of segmented nucleus regions according to their extracted features.

## 1.1 Problem Definition and Motivation

In order to color the Pap smear slides, a dye of Hematoxylin and Eosin is used for staining the nucleus and cytoplasm. Basically Hematoxylin stains the nucleus and combination of Hematoxylin with Eosin stains the cytoplasm. After the staining procedure we get a Pap smear slide where nuclei and cytoplasm parts are colored with the tones of red and blue which makes analyzing the cells on the slide easier. We present an example cell from a Pap smear slide with its background, cytoplasm and nucleus after the staining procedure in Figure 1.1.

There are thousands of cells in a typical single Pap smear slide. The slides are scanned with different magnification levels using a microscope by cytotechnicians in order to detect the abnormal cells in the slides. These main magnification levels are 10x, 20x, 40x and 100x. Each of these magnification levels has its own

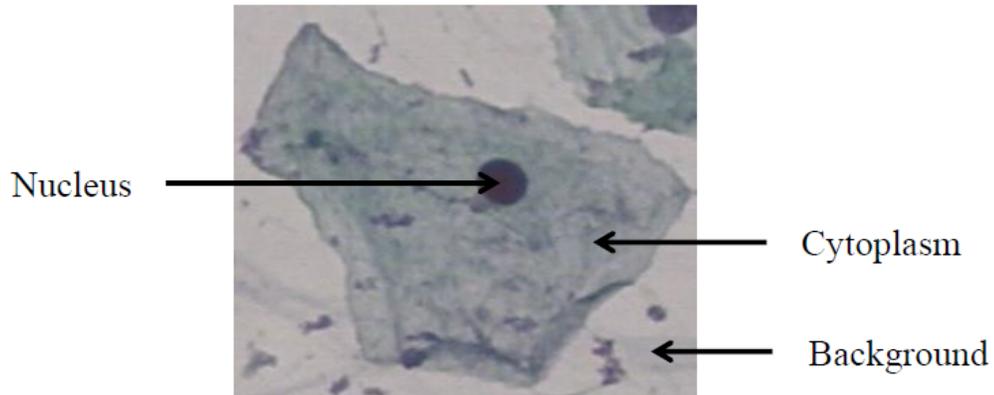


Figure 1.1: An example cell from a Pap smear slide with its background, cytoplasm and nucleus after the staining procedure.

task like identification of background, close up view for overlapped, occluded and grouped cells and examining the size, color, shape and texture of a single cell in details.

There are several difficulties that are associated with the Pap smear test. As a result of traditional Pap smear staining test technique, we have high number of cells including overlapped, occluded and grouped cells. Identifying these occluded and overlapped cells requires different settings in terms of magnification and focus. The other problem is that, in addition to nucleus and cytoplasm there are inflammations and other microorganisms in the Pap smear slides (see Figure 1.2). Also, staining of cells in a Pap smear slide is not homogeneous and the contrast between nucleus and cytoplasm is usually low. Figure 1.2 illustrates these main problems for a 20x magnification Pap smear image.

The first step of diagnosing cervical cancer is to classify the cells in a Pap smear slide as normal and abnormal. The categorization in [4] further divides normal cells into three subcategories called *Superficial*, *Intermediate* and *Columnar*. Table 1.1 summarizes normal degree cells with their main characteristics.

Abnormal cells have four different categories according to their cancer risk. In order from lower risk to higher risk, they are *Mild dysplasia*, *Moderate dysplasia*, *Severe dysplasia* and *Carcinoma in situ*. When cancer risk is increasing, the nuclei of the cells is getting larger, darker, also nucleus is more deformed and the

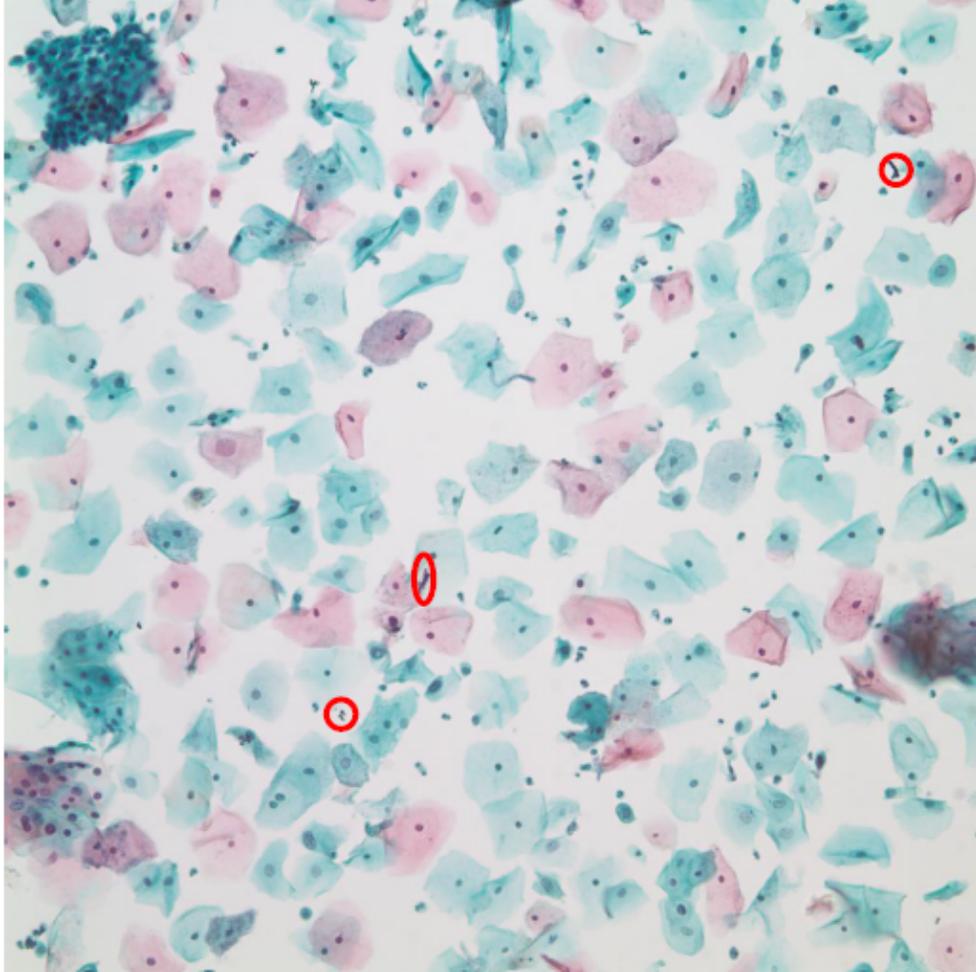


Figure 1.2: An example 20x magnification Pap smear image with inconsistent staining, poor contrast, grouped, occluded and overlapped cells. The red circles depict inflammations and other microorganisms.

ratio of nucleus and cytoplasm area is higher (see Table 1.2 ). As it could be observed from the Table 1.1 and Table 1.2, the given precancerous and cancerous cells differ in their morphological characteristics like size, color, shape and texture of both nucleus and cytoplasm.

Most of the work in the literature works on individual cells where the problem is simple contour finding of nucleus and cytoplasm. However, in real world settings, we have much more complex cell structures including occluded, overlapped and grouped cells and it is impossible to have all these cells isolated from each other in Pap smear slides. So, in order to present effective and realistic solutions,

we work on real world dataset which includes all these mentioned difficulties and problems above.

In this thesis we present a study to segment and order the nuclei of the cells according to their abnormality degree. The presented approach in this study is motivated by the way which is used by cytopathologists to detect the abnormal cells. At the first stage, the pathology experts use lower level magnifications to select potential cells in particular parts of the Pap smear slide. Then, they switch to higher magnification levels to have closer look at this part of the Pap smear slide in order to observe the characteristics of the cells like size, color, shape and texture. They mostly use 10x magnification and/or 20x magnification as low level magnification; and 40x magnification and/or 100x magnification as high level magnification. Since it is even very difficult for an expert to differentiate the boundaries of grouped cells which overlap and occlude each other, they mainly consider the nuclei of the cells while making their decisions.

Based on these facts and inspired by human way of examination of cervical cells, in this study we focus on only the segmentation and ordering of nuclei regions in Pap smear slide images. Therefore, we first segment the Pap smear images at 20x magnification, and following segmentation, we eliminate some of the segmented regions in order to obtain only the nucleus regions by using four different extracted features from the segmented regions of 20x magnification Pap smear image. Then we segment the remaining nucleus regions over 40x magnification and extract effective features from 40x magnification. We extract 15 different features and apply six different ordering algorithms to rank the nucleus regions according to their dysplasia degree. We test the ordering algorithms with different combination of features. In our study, we use a non-parametric hierarchical segmentation algorithm and we sort the segmented nucleus regions by applying different ordering algorithms in an unsupervised way. Figure 1.3 summarizes the steps of our presented system.

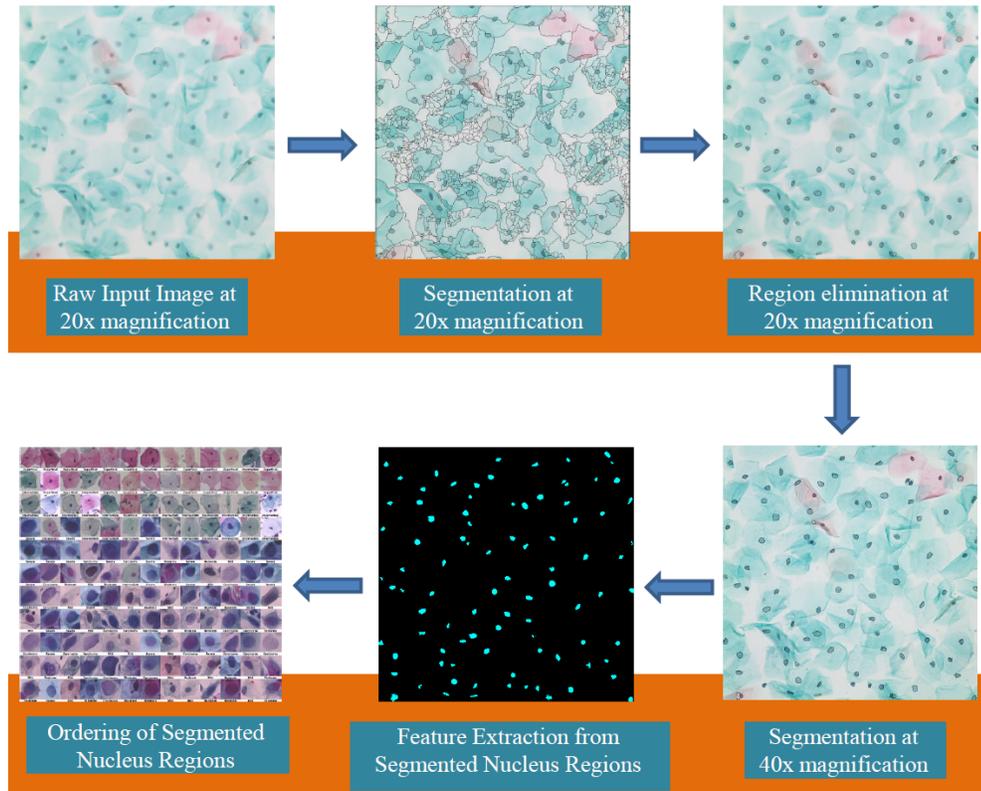


Figure 1.3: Main steps of the proposed automatic segmentation and ordering procedure for the cells in Pap smear Images.

## 1.2 Data Set

We have two different data sets, namely Hacettepe and Herlev data set. Below we give details of these data sets.

### 1.2.1 Hacettepe Data Set

The Hacettepe data set was collected at the Department of Pathology at the Hacettepe University Hospital under the supervision of Dr. Sevgen Önder from Hacettepe University. To capture images of Pap smear test slides, we used a microscope connected to a digital camera.

Our dataset includes 252 images from three different patients' Pap smear test

slides. We captured images of same areas in Pap smear test slides at three different magnification levels . There are 84 images from each of these three Pap smear slides; four of them at 10x magnification, 16 of them at 20x magnification and 64 of them at 40x magnification. Then we generated image triplets for seven different areas in which there are three images for 10x magnification, 20x magnification and 40x magnification levels. Figure 1.4 illustrates one of these triplets as an example. As it can be seen from the figure, when magnification level increases, we see more details of cells in the Pap smear images. However, in our study 10x magnification images are not used, only 20x magnification and 40x magnification image pairs are used for segmentation.

### **1.2.2 Herlev Data Set**

Herlev data set was collected by the Department of Pathology at Herlev University Hospital and the Department of Automation at Technical University of Denmark. In this data set there are 917 images of Pap smear cells [4]. Each image includes only one cell with its nucleus, cytoplasm and background, and each of these cells are manually classified into one of the seven classes by doctors as presented in Table 1.1 and Table 1.2. Since we have ground truth order of these cells according to abnormality degree, we use Herlev data set to show our ordering results.

## **1.3 Contributions**

In this study we aim to obtain an order of segmented nucleus regions according to their abnormality degree. Once we get this ordered list of nuclei, it will be enough for doctors and cytotechnicians to focus on abnormal cells in the ordered nuclei list. In this way the diagnosing process will be more efficient and take less time by investigating only the candidate nucleus regions.

Ordering algorithms are applied on segmented nucleus regions; therefore the

most important step is accurate segmentation of nucleus regions. As mentioned and illustrated previously, there many grouped, overlapped and occluded cells in the Pap smear images; therefore in this study we aim only the segmentation of nucleus regions in the given Pap smear images in order to have realistic results. For this purpose, we follow human’s approach to the segmentation problem where we first segment Pap smear images at low level magnification, which is 20x magnification in this study, and we only choose the regions which are considered as nucleus regions. After we obtain the nucleus regions from 20x segmentation, we switch to 40x magnification to extract good quality features in terms of morphological properties of nucleus regions such as size, color, shape and texture. Finally we apply ordering algorithms using these extracted features to order the segmented nucleus regions.

Differently from most of the studies in the literature we work on real world data set collected from the Department of Pathology at Hacettepe University which includes grouped, overlapped and occluded cells. Moreover the captured Pap smear images have inconsistent staining and poor contrast between cytoplasm and nuclei. Our segmentation method has human inspired approach and it is a an unsupervised algorithm. Both segmentation and ordering process are unsupervised processes and they do not require learning step as well training and test sets.

This thesis is organized as follows. Chapter 2 gives a brief summary of previous studies related to segmentation and classification/ordering of cells, especially for the Pap smear images. In Chapter 3, we explain our segmentation method for 20x magnification and 40x magnification images in detail. In Chapter 4, we first give details of our extracted features from 40x magnification images and then we describe our distance calculation methods to obtain distance matrixes. In Chapter 5 we explain ordering algorithms which are used to order the segmented nucleus regions, and finally in Chapter 6 we present our experimental results for both segmentation and ordering algorithms.

Table 1.1: Normal Cells

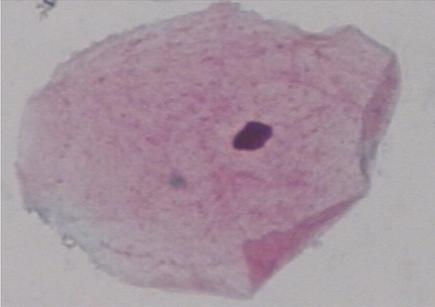
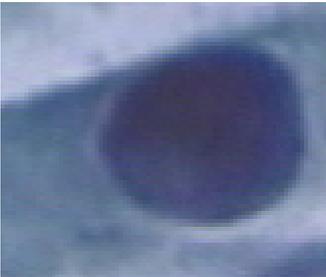
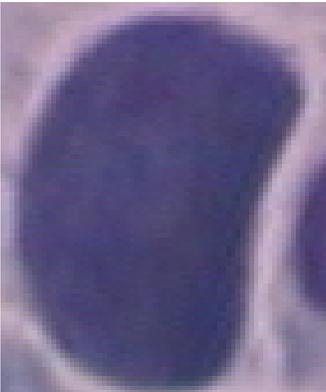
Figures	Characteristics
	<p><b>Superficial Cell;</b> Oval shape, Very small size nucleus, Small ratio of nucleus/cytoplasm.</p>
	<p><b>Intermediate Cell;</b> Round shape, Small size nucleus, Small ratio of nucleus/cytoplasm.</p>
	<p><b>Columnar Cell;</b> Column-like shape, Larger size nucleus, Medium ratio of nucleus/cytoplasm.</p>

Table 1.2: Abnormal Cells

Figures	Characteristics
	<p><b>Mild Dysplasia;</b> Light color nucleus, Large size nucleus, Medium ratio of nucleus/cytoplasm.</p>
	<p><b>Moderate Dysplasia;</b> Dark color nucleus, Large size nucleus, Large ratio of nucleus/cytoplasm.</p>
	<p><b>Severe Dysplasia;</b> Dark color nucleus, Large size nucleus, Deformed nucleus Very large ratio of nucleus/cytoplasm.</p>
	<p><b>Carcinoma in situ;</b> Dark color nucleus, Large size nucleus, Deformed nucleus Very large ratio of nucleus/cytoplasm.</p>

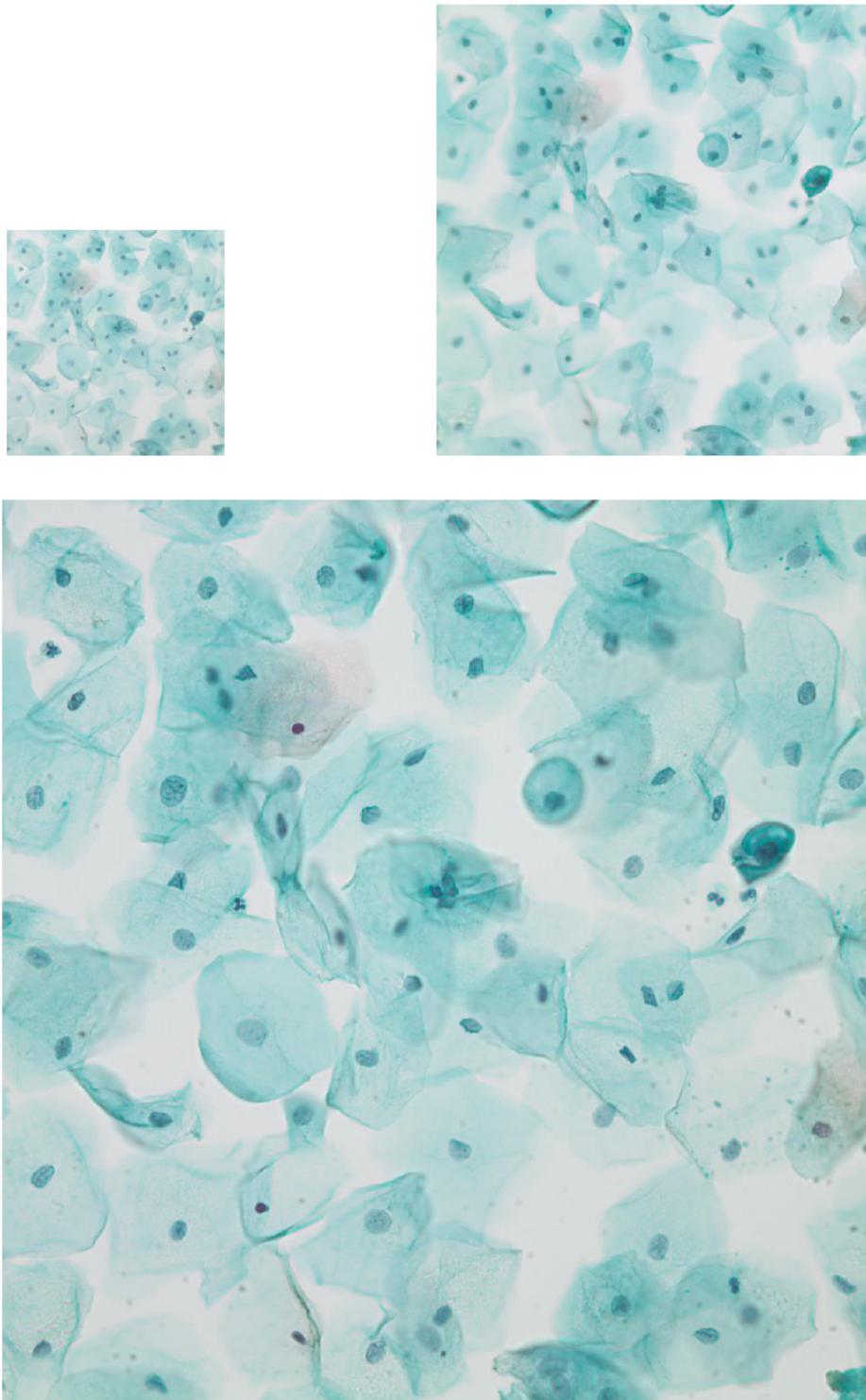


Figure 1.4: Three Pap smear images of the same area with the size of  $512 \times 512$ ,  $1024 \times 1024$  and  $2048 \times 2048$  respectively correspond to 10x, 20x and 40x magnification levels.

# Chapter 2

## Related Work

In this section, we present a brief survey of some previous works related to segmentation.

Automatic thresholding, morphological operations and active contour models are the popular methods that are used for segmentation of the cells in Pap smear images. In the literature, much of the works consider single cells as an input to segment only a nucleus and its cytoplasm. However, in real world settings we have grouped cells that overlap and occlude each other.

As one of the popular methods, automatic thresholding [5, 6] could give good results for the isolated single cells. In case of enough contrast, active contour based methods [7, 8, 9, 10, 11] are successful to extract better localized nuclei boundaries, but they are very sensitive to parameters and initializing process. Watershed algorithms are another common approach. Watershed-based methods [9, 12, 13] are more successful to segment multiple cells in the given images, but they require preprocessing, especially for selecting markers. Using shape priors together with active contour methods could be a solution for the overlapping and occluded cells [14]. However, there are still unsolved problems such as the number and location of cells related to this approach; and also to define a prior shape for the overlapped and occluded cells is another main problem. In the following, we describe some of the selected works in details.

In study [5], authors only consider single cells and find contours of both nucleus and cytoplasm. They first do pre-processing to enhance the edges of nucleus and cytoplasm, then they apply automatic thresholding to obtain nucleus and cytoplasm regions. This method is valid the only an isolated single cell; in case of overlapped or occluded cells it fails.

Bamford and Lovell [7] use a viterbi search-based dual active contour algorithm where they estimate their active contour model in a dynamic way. This approach is based on the contrast between nucleus and cytoplasm. They mark a point inside the nucleus, considering nucleus is darker than its cytoplasm. For this purpose, they first reduce the search space and find the nucleus contour by minimizing some cost function. However, this approach tends to fail in case of inappropriately arranged parameter for the global minimum.

Dagher and Tom [8] present a new approach to the segmentation problem for blood and corneal cells. They basically combine watershed algorithm and the active contour model. They prepare the images by removing noise and then they use down sampled watershed segmentation result to initialize the snake contours for the nuclei. The difficulty of this approach is finding initial contours of nucleus and also active contour models require many parameters to tune.

Huang and Lai [9] aim to find approximate segmentation for liver cells in the biopsy images by eliminating non-nucleus regions in a heuristic way. For the segmentation, first they apply marker-based watershed algorithm to find approximate boundaries of nucleus regions, then they use snake model to refine these boundaries. However, finding marker for all nucleus regions is nearly impossible due to overlapping and occluded cells in Pap smear images.

Harandi et al. [10] present a segmentation method for the Thin Prep slide images which uses active contour algorithm to extract the cell boundaries in cell groups. As a similar approach to our study, they use two different resolution levels. They use lower resolution images to find the regions of interest, and higher resolution images for segmentation. However, in their study they work on specific parts of slide images where there is no inflammation and other microorganisms.

Li et al.[11] roughly segment an image into nucleus, cytoplasm, and background regions applying k-means clustering, and then they use snake algorithm to improve segmentation results for nucleus and cytoplasm. k-means clustering is also preferred by Tsai et al.[6] as a thresholding method to extract the cells from background.

Plissiti et al.[12] first detect nuclei centroids and use the detected centroids as markers for the watershed segmentation to obtain the boundaries of nucleus regions. Then they extract shape, texture, and intensity features and obtain nucleus regions by using a binary SVM classifier with the features.

The study presented by Wu et al. [15], also aims to segment a single cell image. The method uses prior information of nucleus such as the shape, size and contrast between nucleus and its cytoplasm. They use a parametric cost function to extract the boundary of the given single cell by assuming that the nucleus is in an elliptical shape. This approach also is not suitable for the segmentation of Pap smear images where there are many grouped, overlapped and occluded cells.

In the work presented by Walker et al.[16] they segment nucleus regions by removing cytoplasm regions using morphological closing operation. Following this, they apply morphological opening operation to correct the obtained nucleus regions. Since they use global thresholding to remove cytoplasmic parts, it tends to fail depending on image structure.

Shah [17] calculates the approximate cell locations at the first step based on a clustering approach. In the second step he uses an ellipse shape as prior information to find the final cell locations. This method shows good performance for the Pap smear images taken at lower magnifications.

In [18] authors segment a single-cell image into nucleus, cytoplasm and background region by using the fuzzy C-means (FCM) clustering technique.

The study in [19] aims to segment the individual cytoplasm and nuclei in a group of overlapping cervical cells. In this study authors first specify single cells and grouped cells together with their nuclei, then they perform a joint level set

optimization on these specified nuclei and cytoplasm pairs. This optimization basically includes a set of restrictions in terms of the length and area of each cell, a prior on cell shape and the amount of cell overlap.

In study [20], authors propose a multi-scale watershed-based method to segment nerve cell nuclei. They apply watershed segmentation algorithm at different scales and select a set of regions by thresholding regions' features.

Among these previous works, there are a few studies which take account real world data set with the main challenges of grouped, overlapped and occluded cells, and poor contrast. In [21] Gençtav and Aksoy present a non-parametrical segmentation algorithm to segment Pap smear images. For this purpose they first extract background using an automatic threshold, and then they apply their hierarchical segmentation algorithm to detect nucleus and cytoplasm regions.

Since the aim of segmentation is to detect the abnormal cells in the images, following segmentation, many studies classify segmented cells. Huang and Lai [9] classify hepatocellular carcinoma cells, which is a common type of liver cancer, in biopsy images using an SVM-based graph classifier. In order to classify cervical cells, Walker et al. [16] extract textural features from co-occurrence matrix and classify the cervical cells according to these features by using a quadratic Bayesian classifier. Neural networks are used to classify blood cells by Theera-Umpon [22] as a classification method. In [23] authors use a hierarchical multiple classifier with more than 300 features to classify the segmented cells. A pixel-based classification method is used by Zhang and Liu [24] with 4,000 multispectral features.

Most of these works explained above classify cells into two classes namely normal and abnormal. Different from these works Marinakis et al. [25] consider this problem as multiclass classification where the number of classes is seven. They extract 20 features computed from nucleus and cytoplasm regions and apply a genetic algorithm to select features, and then classify the regions by using a nearest neighbor classifier. However, compared to binary classification results, they obtain less successful results.

The studies show that when the classification problem of cells is considered as

normal and abnormal cell labeling, we obtain higher accuracy scores. However, these results are obtained with a limited number of instances in datasets and mostly they are synthetically prepared and controlled data sets which do not include the main challenges such as grouped and overlapped cells. The other point is that classification requires a large dataset where there should be enough samples for each class to be used in the training procedure. At this point, as a real world problem, we have two main challenging facts related to classification. The first one, we have seven different categories for the cells, which make classification even harder. The second one, we have imbalanced data in which among hundreds of cells, the frequency of observing abnormal cells is very small; so it is nearly impossible to have a sufficient number of cells for each class.

Based on these facts and in order to present realistic solutions, we approach this problem as an ordering problem rather than a classification problem. With this approach we aim to get an ordered list of nuclei in the Pap smear in which normal cells are conglomerated at one end and abnormal cells are conglomerated at the other end.

# Chapter 3

## Segmentation

In the segmentation step, we aim to obtain an accurate segmentation of nucleus regions in the most correct way. However, segmentation of cell nucleus is a difficult task due to the reasons beyond our control. One main problem is the traditional staining techniques that are used to color cervical cells on a Pap smear test slide with the tones of blue and red colors. These traditional staining techniques cause inhomogeneity in the slide and also inconsistency between different slides. In addition to inconsistent staining, grouped cells usually overlap or occlude each other. Even manually, it is not easy to differentiate boundaries of overlapping cells. Figure 3.1 illustrates two different Pap smear images which are at 20x and 40x magnification with these mentioned problems. Therefore, segmentation of Pap smear test images is still a challenge due to inconsistent staining, poor contrast and overlapping cells.

In Figure 3.1, we show two Pap smear images at 20x and 40x magnification of same slide area. In the figure, the Pap smear images have size of 1024 and 2048 in each dimension corresponding to 20x magnification and 40x magnification respectively. To see the main differences between 20x and 40x magnification in Figure 3.2, we show the close up views of a small Pap smear region at 20x and 40x magnification. As it could be observed from this figure, the 40x magnification image has more detailed texture but worse contrast compared to 20x magnification image. Following our segmentation, we rank the segmented nucleus

regions according to their extracted features. Contrast is an important property for segmentation and detailed texture is an important property for feature extraction. With these facts we end up with a tradeoff where it is better to use 20x magnification images for segmentation and 40x magnification images for feature extraction.

To overcome this tradeoff, we propose a two-phase approach to segmentation problem. The first phase is the segmentation of Pap smear images at 20x magnification. Following segmentation, we eliminate some of these segmented regions which are potentially not nucleus. The final phase is the segmentation of remaining 20x magnification nucleus regions over 40x magnification images. The details of each step are explained in the following sections.

### 3.1 Segmentation Method

Pap smear images have three main regions which are background, cytoplasm and nucleus. However, because of the factors like overlapped cells, inconsistent staining and poor contrast in Pap smear images, it is nearly impossible to segment cytoplasm of each nucleus accurately. Therefore, to have realistic results, in our segmentation we focus on obtaining only the nucleus regions in the most correct way.

In their work, Gençtav and Aksoy [21] present a study for segmentation and classification of cervical cells. Basically they first extract background regions using a threshold value to obtain cells. Then, they segment the remaining cell regions using a hierarchical segmentation algorithm. Finally they classify the segmented regions as nucleus or cytoplasm region. In our study we aim to segment only nucleus regions by following and modifying their proposed segmentation method. In this section we give a brief summary of this segmentation algorithm.

The algorithm developed by Gençtav and Aksoy [21] is a parameter free algorithm and it basically uses the spectral, shape and gradient information of the Pap smear images. In their study they first extract the background region which

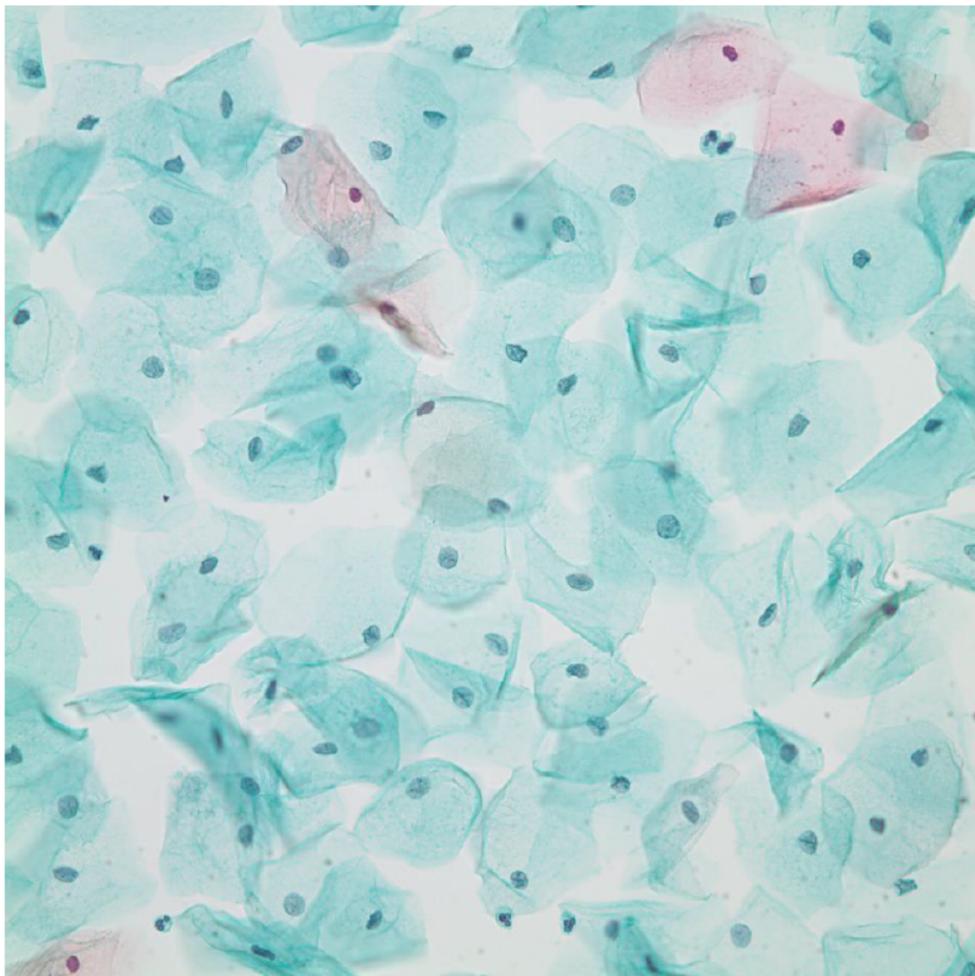
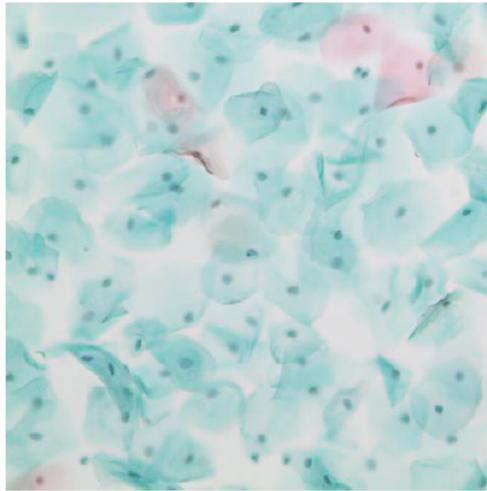


Figure 3.1: 20x (1st row) and 40x (2nd row) magnification Pap smear images with inconsistent staining, poor contrast and overlapping cells.

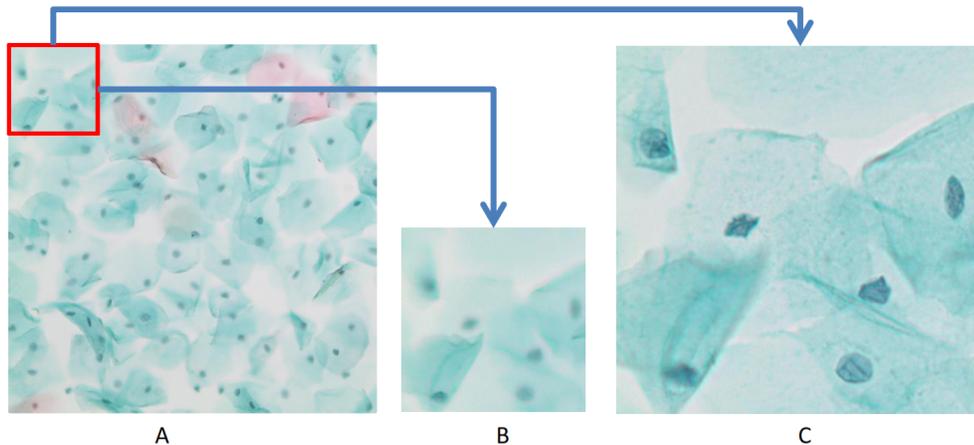


Figure 3.2: The original image at 20x magnification (A), close up view of the red rectangle at 20x magnification which has more contrast (B) and the corresponding image at 40x magnification which has more detail (C).

is the region that does not include any cytological structures and it has fully white pixels. For this purpose, they transform the Pap smear images from RGB color space to the Lab color space. After this transform, they distinguish background from cell regions by using L channel of the Lab color space. In the Lab color space, the L channel corresponds to brightness of the image. As a final step of background extraction, they use minimum error thresholding to determine the threshold value which distinguishes background and cell regions from each other.

Then, following background extraction, they segment the remaining cell regions into the areas of nucleus and cytoplasm. The proposed segmentation algorithm in [21] is based on the work of Akçay and Aksoy [26] where a segmentation method was developed to detect geospatial objects like buildings, roads, etc. automatically. They use the neighborhood, spectral and morphological information and apply morphological opening and closing operations to extract the candidate regions. Later they build a hierarchical tree from the extracted regions and select the most meaningful regions in that tree. To select the meaningful regions, they optimize spectral homogeneity and neighborhood connectivity measure where spectral homogeneity is the variances of multi-spectral features and neighborhood connectivity is the sizes of connected components.

Since Pap smear images have different image structure and objects compared

to remotely sensed images, in [21], the candidate regions are extracted by applying watershed segmentation to h-minima transforms of the image gradient instead of using morphological opening and closing operations. The watershed segmentation algorithm is considered to be one of the effective segmentation methods which does not require any prior information about the segment number in the image. The most important characteristic of this segmentation method is that it models local contrast differences using magnitude of image gradient. Relative contrast between nucleus, cytoplasm and background plays an important role in our segmentation problem especially on identifying nucleus regions. Thus watershed segmentation fits as a suitable solution to extract the candidate regions.

There are many different algorithms to compute watersheds. However, they mostly suffer from over-segmentation when they are computed from raw image gradient. To overcome this problem Gençtav and Aksoy [21] use a multi-scale approach to get accurate segmentation results over Pap smear images. They generate a hierarchical partitioning of cell regions with the dynamics which are related to regional minima of image gradient. Here a regional minimum is formed from a group of neighboring pixels with the same value  $x$  where the pixels on its external boundary have a value greater than  $x$ .

As a result of the multi-scale watershed segmentation algorithm, they obtain a set of nested partitions of a cell region. Later, similarly to [26], they build a hierarchical tree from the multi-scale partitions of a cell region and select the most meaningful segments among different levels of the tree. However, again because of different image structure of Pap smear Images, in [21] differently calculated homogeneity and circularity measures are optimized for the meaningful region selection step. Here nucleus regions are the meaningful regions and it is easier to differentiate them by using their appearance, i.e., their homogeneity and shape features. Therefore, after small segmented regions in the lower levels are merged to form nucleus, the aim is to obtain homogeneous and circular nucleus regions at some higher level. The full formed nucleus regions in the most homogeneous and circular way are the segments we want to obtain. These nucleus regions may stay the same during some number of levels until they merge with their surrounding segments of cytoplasm.

## 3.2 Segmentation at 20x Magnification

In this section, our goal is to segment the cell regions of 20x magnification images and obtain correctly segmented nucleus areas. For this purpose we apply the described segmentation algorithm in the previous section to 20x magnification Pap smear images. Differently from this algorithm we do not extract background regions. As Figure 3.1 shows in our Pap smear images there is not a clear background region that is full of white pixels. Therefore, we do not have enough contrast to distinguish cell regions from background. Based on this fact, we directly apply the segmentation algorithm proposed in[21] to the Pap smear images. Finally we obtain a segmentation map of segmented regions of background, cytoplasm and nucleus where the selected regions are numbered starting from 2 while 1 values represent the background.

However, when we directly apply the algorithm, we have over segmented segmentation result (see Figure 3.3). As it could be seen from the figure, especially background and cytoplasm parts are over segmented. To avoid this case, after segmentation of 20x magnification Pap smear images, we eliminate some regions potentially not nucleus. For this purpose, following the segmentation, we extract four different features for each region which are namely mean intensity, size, circularity and homogeneity. We select only potential nucleus regions by eliminating rest of the regions according to experimentally determined threshold values of these extracted features. These features and their threshold values are discussed below. Figure 3.4 shows each step of the segmentation for a 20x magnification Pap smear image.

### Extracted Features from 20x Magnification Pap smear Images

After applying the automatic segmentation method, a set of features is extracted from each segmented region of 20x magnification Pap smear images. At 20x magnification segmentation step, our goal is to obtain only the nucleus regions by eliminating the rest of the regions those are not nucleus regions. Therefore, we need features to characterize and distinguish nucleus regions from cytoplasm and background regions. Figure 3.5(a) shows an example area of a segmented

20x magnification Pap smear image. As it could be seen from the figure, compared to cytoplasm and background regions, nucleus regions are darker, circular and more homogenous. Based on these criteria we extract four different features from the segmented regions, which are mean intensity, size, circularity and homogeneity. Then, we select only nucleus regions according to experimentally fixed threshold values of these features. To calculate the threshold values of these features, we use three Pap smear images at 20x magnification which are included in our dataset. The threshold values are determined qualitatively based on the experiments which are done on these three Pap smear images.

- **Mean Intensity** feature corresponds to normalized L channel values of Lab color space that are in the range between 0 and 1. However, since Pap smear slides are colored with tones of blue and red, it differs between 0 and 0.4. We experimentally fix threshold value for this feature to 0.13. The regions whose mean intensity is less than 0.13 are eliminated.
- **Size** feature of a region is the total number of pixels in that segmented region. We have two different experimentally fixed threshold values which are respectively 120 and 1060. The value 120 is used to eliminate very small regions while 1060 value is used to eliminate very large regions like background regions.
- **Circularity** feature of each region is calculated as

$$f_{circ} = \frac{4\pi A}{P^2} \quad (3.1)$$

where  $A$  and  $P$  is the area and the perimeter of a region respectively. Since the perimeter of a 1-pixel size region is 0, the circularity of regions is between 0 and 1 for the regions whose size is larger than 1 pixel; the circularity value 1 represents a perfect circular region. The regions with the circularity value less than 0.62 is eliminated. The value 0.62 is determined experimentally.

- **Homogeneity** feature of each node in the hierarchical tree is calculated based on spectral similarity of the region to its parent node by using the F-statistic. In linear regression, F-statistic is used to test the significance

of the variances of two populations. In our problem, F-statistics is used to measure the correlation between the means of two distributions concerning their pooled variance at different levels of the hierarchical tree. According to formula presented in [21], we calculate the homogeneity of a region as follows

$$F(R_1, R_2) = \frac{(n_1 + n_2 - 2)n_1n_2}{n_1 + n_2} \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} \quad (3.2)$$

where  $R_1$  is a node in the hierarchical tree and  $R_2$  is its parent node.  $n_i$ ,  $m_i$  and  $s_i^2$  indicate the number of pixels, the mean of the pixels and the scatter of the pixels for  $R_i$  respectively, where  $i = 1, 2$ .

The threshold value for this feature is set to 20675 experimentally; so that the regions whose homogeneity value is less than this value are eliminated.

The regions which satisfy the threshold criteria for each feature are considered as nucleus regions. Figure 3.5(b) illustrates region elimination result of Figure 3.5(a) according to these threshold values of features.

After we select the regions which are considered as nucleus, we extract each of these regions from the original 20x magnification Pap smear image as an individual image by adding 3 pixels margin to their bounding box position. Each of these individual 20x magnification nucleus regions is used as a template for the segmentation step of 40x magnification Pap smear images.

### 3.3 Segmentation at 40x Magnification

In this section, we explain segmentation of nucleus regions selected from 20x magnification Pap smear images over 40x magnification Pap smear image. Since we manually capture Pap smear images of the same slide area at 20x and 40x magnification, we have registration error due to the drift and optical distortion of the lens. Therefore, calculating corresponding relative positions of extracted 20x magnification regions in 40x magnification images is likely to be inaccurate. Considering this fact, we extract the same regions from original 40x magnification

images by calculating relative positions and adding 35 pixels margin in order to guarantee that the nucleus is in the extracted region. The number 35 is not a significant value for a  $2048 \times 2048$  size 40x magnification Pap smear image. However, it is experimentally determined minimum value to cover the registration error.

Due to the added 35 pixels margin, the extracted regions from 40x magnification images are likely to have unnecessary additional background and cytoplasm part, and as mentioned previously, compared to 20x magnification images, 40x magnification images have much more details (see Figure 3.2). As a consequence of these two facts, the extracted 40x magnification images tend to be over-segmented. Therefore, additional background and cytoplasm parts should be removed to avoid poor segmentation results while saving the image part containing the nucleus. As a possible solution to this problem we attempt to apply template matching between extracted pairs of 20x and 40x magnification Pap smear images. As mentioned in the previous section we extract segmented nucleus regions from 20x magnification images by adding three pixels margin. In this way we obtain nearly a perfect template where nucleus is centered. After we scale 40x magnification nucleus region images by a factor of 0.5, we apply template matching over the extracted and scaled 40x magnification regions by using corresponding extracted 20x magnification regions as a rectangular template  $T$  where the nucleus is in the center. We use sum of squared difference (SSD) as a template matching method which is formulized as

$$\text{SSD}(x, y) = \sum_{x', y'} (T(x', y') - I(x + x', y + y'))^2 \quad (3.3)$$

where  $T(x', y')$  represents pixel values of template image and  $I(x + x', y + y')$  represents the pixel values of the image patch to compare the given template image over the source image by sliding the template.

After we obtain accurate relative positions from template matching process, we extract final regions from the 40x magnification Pap smear images by adding 5 pixels margin. Finally, we apply the segmentation algorithm on extracted regions. Figure 3.6 summarizes the overall segmentation process.

Depending on texture structure of nucleus regions, the extracted nucleus images from 40x magnification could get segmented into more than one pieces (see Figure 3.7(a)). To solve this undesirable case, we use coarse nucleus boundaries obtained from 20x magnification templates. For this, we first resize the 20x magnification templates by factor two to have the same size as 40x magnification images. Figure 3.7(b) shows the steps to obtain the nucleus boundary from corresponding 20x magnification template for the given nucleus in Figure 3.7(a). Later we merge the segmented regions of the 40x magnification image whose at least 75% area overlap with the coarse nucleus region that is obtained from 20x magnification template. Figure 3.7(c) shows the final segmentation result after merging the regions.

As a final step of 40x magnification segmentation, we overlay the segmented 40x magnification images on the original 40x magnification Pap smear image. Figure 3.8 shows the final segmentation result for 40x magnification Pap smear image.

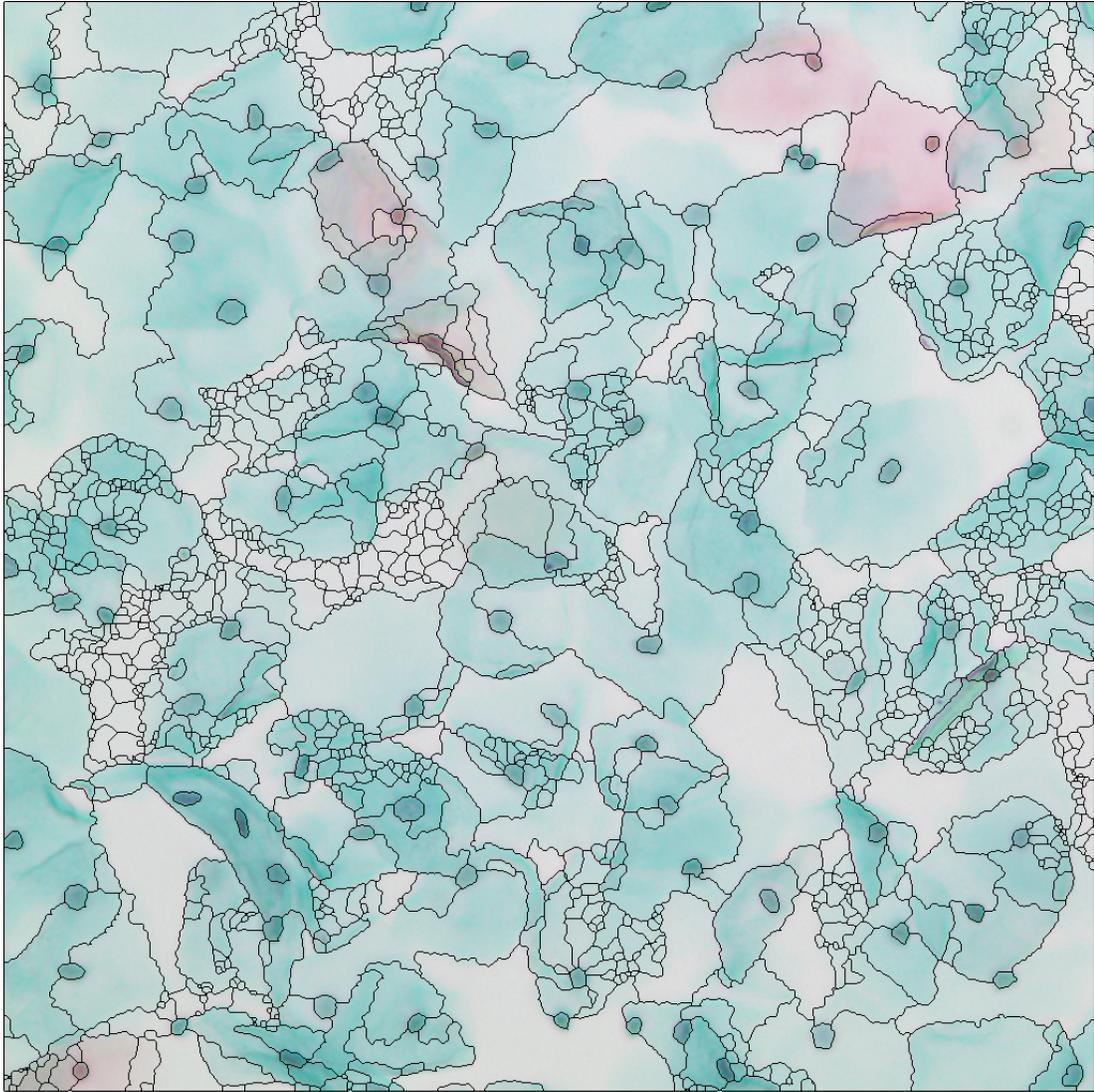


Figure 3.3: Over segmented 20x magnification Pap smear image result when the segmentation algorithm is applied directly.

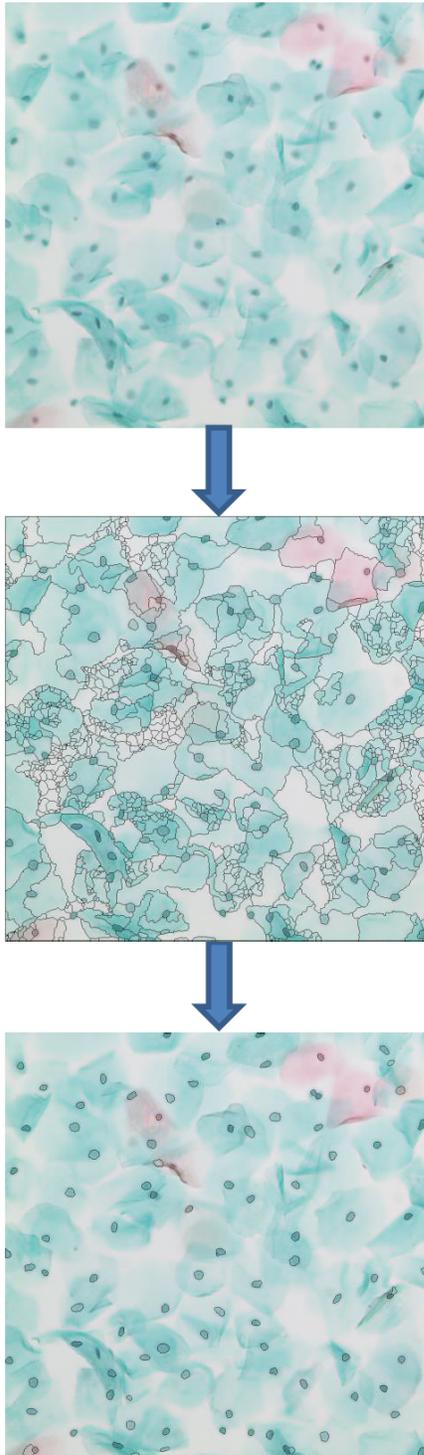
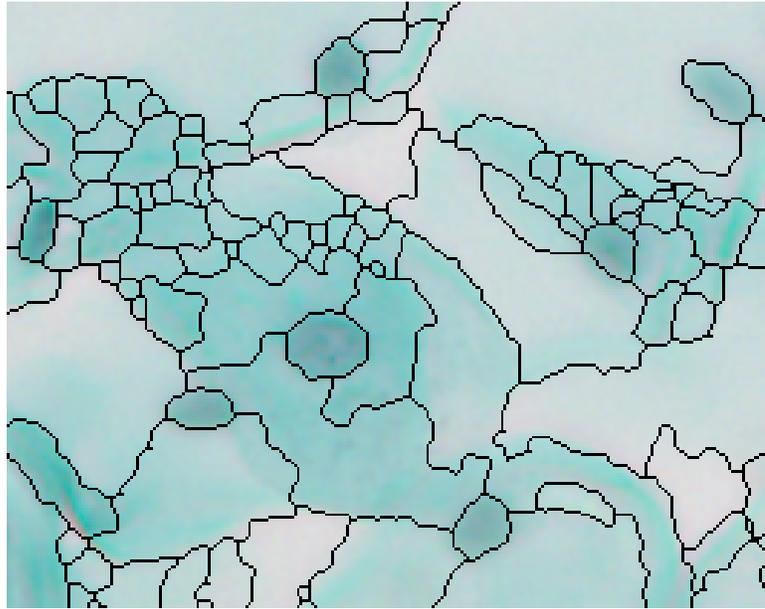
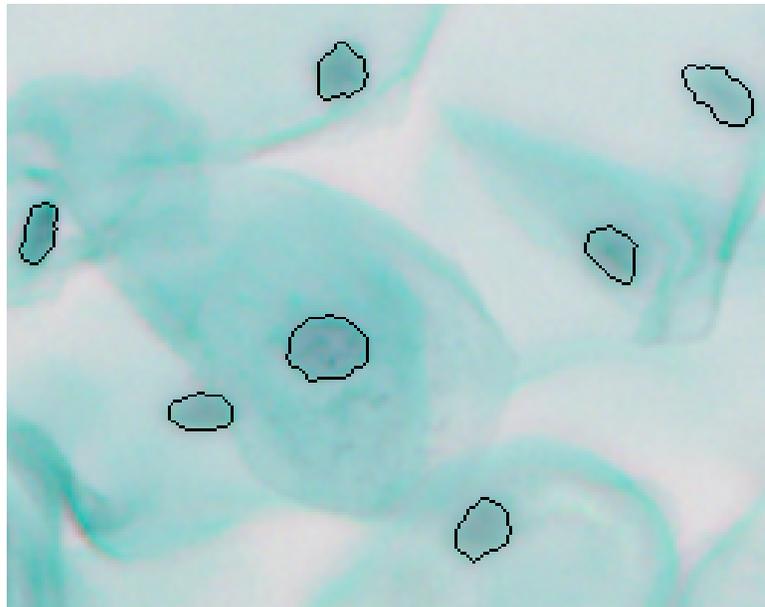


Figure 3.4: Segmentation steps for a 20x magnification Pap smear image. Raw image (1<sup>st</sup> row), the same Pap smear image after the segmentation algorithm is applied (2<sup>nd</sup> row), potential nucleus regions after eliminating the rest of the regions (3<sup>rd</sup> row).



(a)



(b)

Figure 3.5: An example area from 20x magnification segmented image; (a) before region elimination, (b) after region elimination.

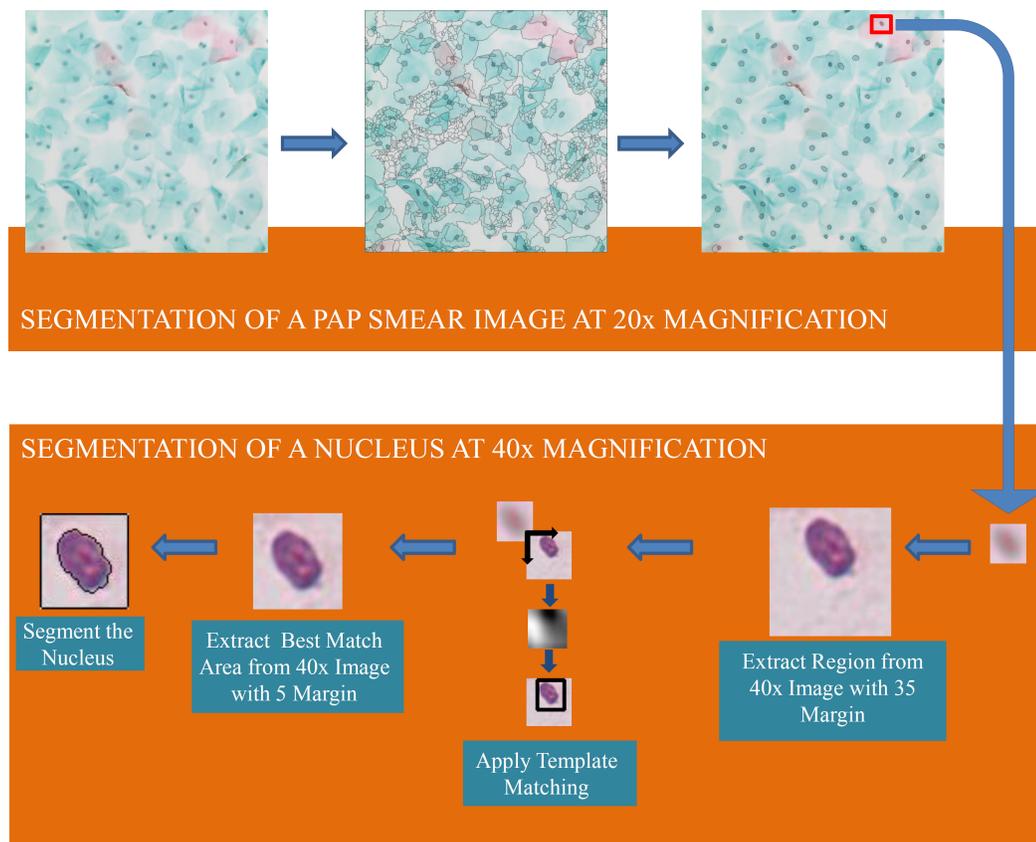


Figure 3.6: The overall process of obtaining a segmented nucleus region from a 40x magnification Pap smear image.

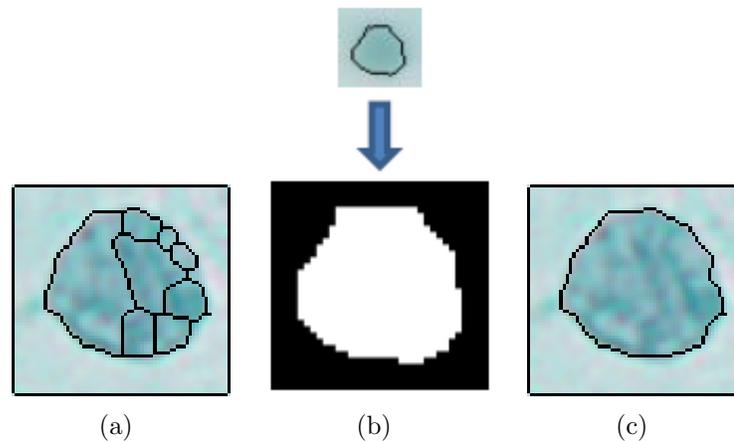


Figure 3.7: Initial segmentation result of the given 40x magnification nucleus image (a), calculated 40x magnification coarse boundary from 20x magnification nucleus template (b), the merged regions whose 75% overlap with the coarse boundary of the nucleus (c).

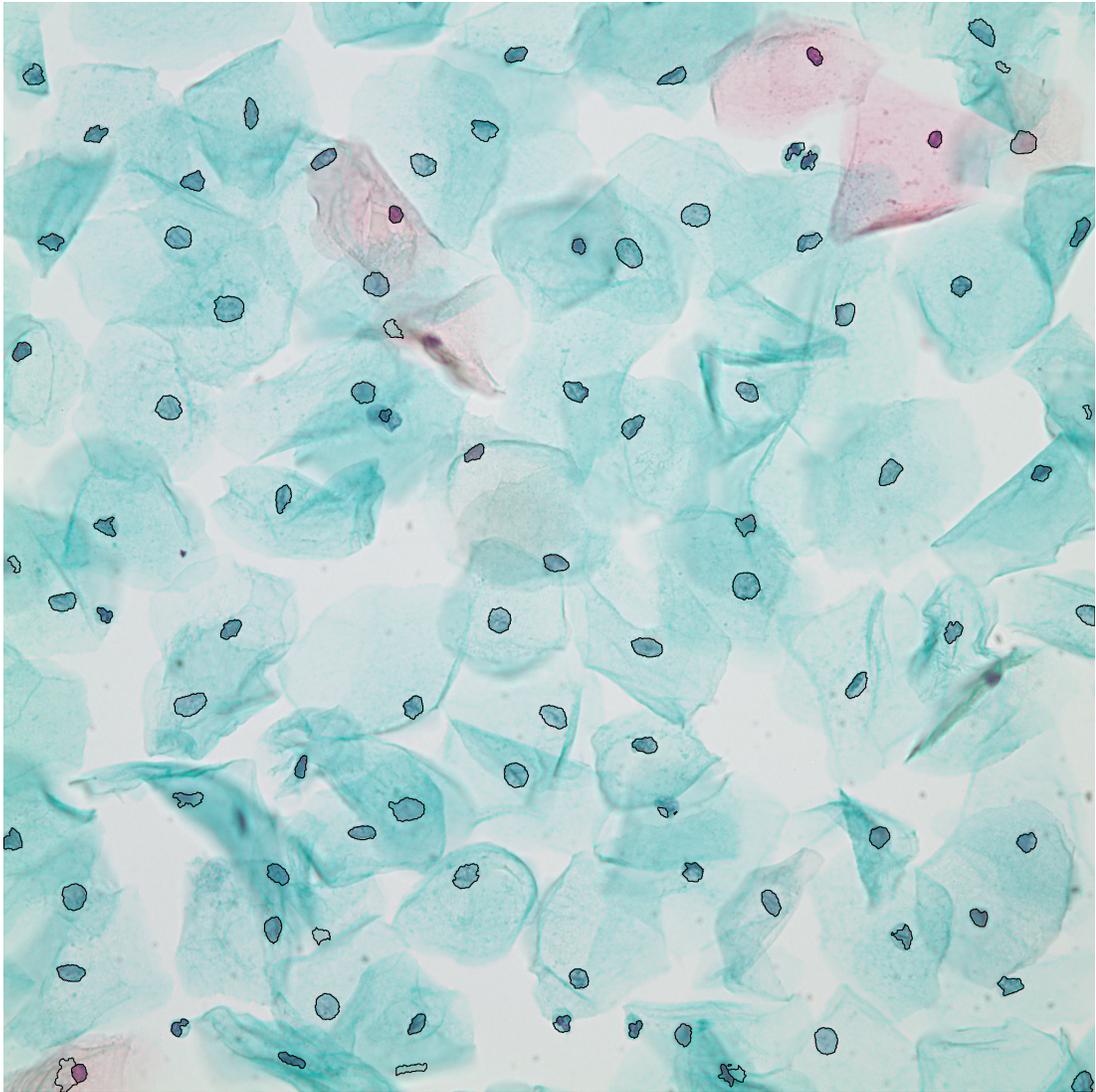


Figure 3.8: Final segmentation result of 40x magnification Pap Smear Image.

# Chapter 4

## Feature Extraction and Distance Measures

In this chapter we first describe and explain the details of our features extracted from 40x magnification. Then, we present our tested methods for combination of features to obtain distance matrix in order to rank nucleus regions.

### 4.1 Feature Extraction

Dysplastic changes and abnormality degree of cervical cells can be determined by analyzing their cytoplasm and nucleus characteristics like size, color, texture and shape. However, as explained in the previous sections in details, it is nearly impossible to segment cytoplasm of each nucleus correctly. Since in Pap smear slide images cells overlap and occlude each other, it is even a difficult task for cyto-technicians and doctors to distinguish cytoplasm of each nucleus. Therefore, in this study we only consider nucleus regions and aim to rank them by using extracted features from only the nucleus regions.

Following segmentation of nucleus regions at 40x magnification, we extract 15 different features from a nucleus region. Eight of these features are defined or

proposed to use by us as follows:

**Contrast** and **homogeneity** features are calculated from the L-channel co-occurrence matrix of a nucleus region. Each element  $(i, j)$  in the co-occurrence matrix represents the number of times that the pixel with value  $i$  occurred horizontally adjacent to a pixel with value  $j$  for four different offsets. At this point, contrast value is the intensity contrast between a pixel and its neighbor over the whole image, so that a constant image contrast value is 0. Contrast feature is calculated as

$$\sum_{i,j} |i - j|^2 p(i, j) \quad (4.1)$$

where  $i$  and  $j$  specify position of an element in the co-occurrence matrix, for row and column respectively; and  $p(i, j)$  is the cell value of the co-occurrence matrix at  $(i, j)$ .

Homogeneity is the value that measures the closeness of the distribution of elements in the normalized L channel co-occurrence matrix to its diagonal. Homogeneity feature is calculated as

$$\sum_{i,j} \frac{p(i, j)}{1 + |i - j|} \quad (4.2)$$

where  $i$  and  $j$  specify position of an element in the co-occurrence matrix, for row and column respectively; and  $p(i, j)$  is the cell value of the co-occurrence matrix at  $(i, j)$ .

**Local binary patterns (LBP)** feature is the special case of the Texture Spectrum model which is proposed in [27] [28]. LBP is one of the efficient texture models in the literature. Basically it labels pixels in the image by thresholding the neighborhood of each pixel in binary way. It has advantages like computational simplicity, suitability for real-time settings and robustness to the variations caused by illumination. In order to extract LBP features, we use the Matlab implementation presented in [29] [30] where a resulting LBP feature contains a rotation-invariant LBP histogram of a nucleus region image in a (8,1) circular neighborhood where 8 pixels are sampled in a circular fashion with 1 pixel radius around a centered pixel.

*Mean intensity of a* and *Mean intensity of b* features correspond to the normalized a and b channel values of Lab color space respectively.

We use the *Mean intensity*, *Size* and *Circularity* features as explained in the previous chapter.

The remaining seven features described below are a subset of the features used in [4] for characterizing cervical cells.

*Nucleus elongation* is the ratio between the shortest diameter and the longest diameter of the segmented nucleus region.

*Nucleus roundness* is the ratio between the nucleus area and the area bounded the circle given by the nucleus longest diameter.

*Nucleus perimeter* is the perimeter length of the nucleus region.

*Nucleus Longest Diameter* is diameter of the smallest circle that circumscribes the nucleus region and calculated as the largest distance between two pixels on the border of the nucleus region.

*Nucleus Shortest Diameter* is diameter of the largest circle that is encircled by the nucleus region.

*Nucleus Maxima* and *Nucleus Minima* is the number of pixels each of which is the maximum/minimum value inside of a  $3 \times 3$  window centered on it.

In the *Chapter 6* we test different combinations of these features in order to determine the optimal number and combination of the features.

## 4.2 Distance Measures

In this section we present our approaches for computing distance matrixes in order to combine multiple features. Since our ordering methods require a distance matrix with positive values as an input for the ranking of the nucleus regions (see

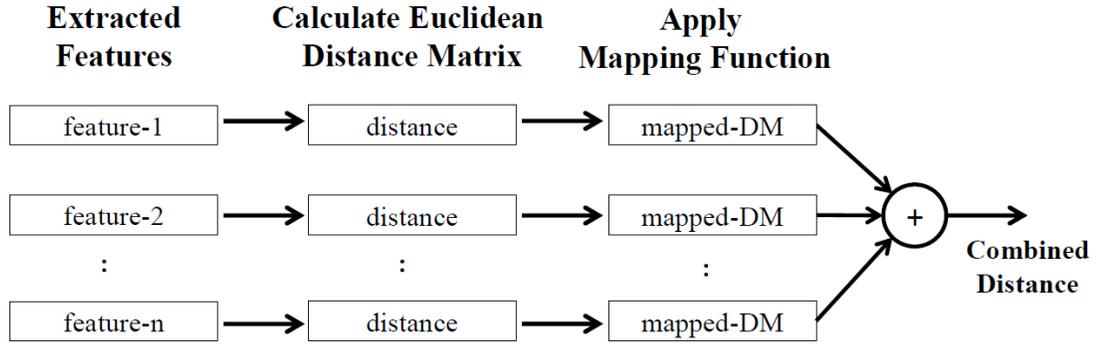


Figure 4.1: An overview of combining distance matrixes obtained from features

*Ordering* chapter for more details), we need to obtain a distance matrix from a set of different combinations of the features.

Figure 4.1 shows an overview of our steps to compute an ultimate distance matrix from multiple features. As it could be seen from the figure, we first compute distance matrixes of each feature using the Euclidean distance metric and we apply standart z-score normalization to each of these distance matrixes as

$$X' = \frac{X - \mu}{\sigma} \quad (4.3)$$

where  $X$  represents one of these distance matrixes,  $\mu$  is the mean of the matrix elements and  $\sigma$  is the standard deviation of the matrix elements;  $X'$  corresponds to the final zscore normalized distance matrix.

After z-score normalization, each element of  $X$  is centered to have mean 0 and scaled to have standard deviation 1. Approximately 95% of the elements of the distance matrix have z-score value between -1 and +1.

Later we map the distance matrices to a new space using different functions. We aim to observe whether these mapping functions is helpful to improve the results. Next, obtained distance matrix of each feature is combined by adding them together. In the last step, if the minimum value of the final distance matrix is a negative number, we shift this matrix to positive zone in a linear way by subtracting the minimum value from the elements of distance matrix.

We use 5 different functions for mapping of distance matrixes. Below we

explain our functions to map and compute a distance matrix of a feature.

**Method1:** No mapping function. We only shift the distance matrix after z-score normalization as

$$DM_n = X' - \min(X'). \quad (4.4)$$

where  $DM_n$  indicates distance matrix of  $n^{\text{th}}$  feature

**Method2:** Log-sigmoid transfer function is the mapping method. Distances matrix elements are calculated and scaled into the range between 0 and 1 as

$$DM_n = \frac{1}{1 + e^{(-X')}}. \quad (4.5)$$

**Method3:** Exponential function is the mapping method. Distances matrix elements are mapped to the interval (0,Inf).

$$DM_n = e^{(-X')}. \quad (4.6)$$

**Method4:** Mapping function is the square root of each element in the distance matrix. Since square root function is valid for the positive values, we first shift the z-score result matrix, and then map this distance matrix by taking the square root of the distance matrix.

$$DM_n = \sqrt{X' - \min(X')} \quad (4.7)$$

**Method5:** Square root is again the mapping function but in a different way. This time we first take the square root of the raw distance matrixes, later we apply z-score normalization and finally shift the distance matrix.

$$\begin{aligned} Y &= f_{zscore}(\sqrt{X}) \\ DM_n &= Y - \min(Y) \end{aligned} \quad (4.8)$$

# Chapter 5

## Ordering

Finding a linear order for the objects of a dataset is a basic and important problem of data analysis and pattern recognition. Ordering algorithms aim to get a sorted list of the objects in a dataset by optimizing specific functions.

In this section, we introduce our ordering problem and methods to order the segmented nucleus regions according to their abnormality degree. Basically we aim to get an ordered list of nucleus regions where they are sorted from normal nuclei to the most abnormal nuclei. In this way cytotechnicians or doctors could save time by skipping normal nucleus regions and focus on only the cancerous nucleus regions.

Classifying cells according to their abnormality degrees is a well-researched problem of medical imaging and many different supervised approaches have been studied for this purpose. However supervised methods require large training sets for the learning phase to classify segmented nucleus regions. Collecting such a large training set is a difficult and challenging task due to previously mentioned facts like overlapping cells, and inconsistent staining. Moreover compared to normal cells, frequency of dysplastic/abnormal cells is quite small; therefore, it is not realistic to collect a balanced, large training dataset which has sufficient number of cells for each class. Training the supervised methods with imbalanced datasets mostly induces biased results. Within this framework, unsupervised ordering

methods are promising as they do not require any learning phase. However, it has two main difficulties which make ordering challenging. The first one is that; we need to get an ordered list with multiple criteria. In our ordering problem, our multiple criteria are different combinations of the extracted features. Secondly, we do not have a reference point to order the nucleus regions. Except these facts, due to the nature of combination and permutation, time complexity could get worse with the size of objects in the dataset and number of dimension which are the features in our case.

### Definition of Ordering

Given  $n$  objects in a dataset which is  $\{O_1, \dots, O_n\}$ , we first compute an  $n \times n$  symmetric dissimilarity matrix  $D$  where  $D(i, j)$  represents the dissimilarity between the  $i^{\text{th}}$  and  $j^{\text{th}}$  objects of the dataset. Later according to a defined optimization function we reorder the dataset by minimizing a loss function or maximizing a merit function as

$$\min Loss(\varphi(D)) \text{ or } \max Merit(\varphi(D)) \quad (5.1)$$

where  $\varphi$  is the defined permutation function in order to reorder the elements of  $D$  by permuting rows and columns at the same time.

In this thesis, in order to sort extracted nucleus regions from 40x magnification Pap smear images, we apply different ordering algorithms on the nucleus features extracted before. For this purpose, we use the R seriation package introduced in [31]. In [31] authors implement different existing algorithms with R project. In the following we first explain the details of the implemented ordering algorithms in [31] and their usage in our dataset.

- **Hierarchical clustering (HC)**

Hierarchical clustering is one of the most popular clustering algorithms used in biological research, especially for the works related to genes [32, 33, 34, 35]. The idea behind the algorithm is producing nested clusters where each of them can be represented as a binary tree. In this binary tree data structure nodes are placed according to their similarities. Even though this

method is more like a clustering approach, still we could use the leaves of produced binary tree as an ordered list.

- **Hierarchical Clustering Reordered by Optimal Leaf Ordering (OLO)**

This ordering algorithm is an extended version of HC (Hierarchical Clustering). The algorithm first performs hierarchical clustering, and then improves the result of hierarchical clustering with optimal leaf ordering approach by minimizing the Hamiltonian path. In graph theory, the Hamiltonian path corresponds to a path which visits each vertex exactly once in an undirected or directed graph. In our ordering problem, the vertexes are our nuclei and the edges of the undirected graph are the distances between two nucleus regions based on the extracted features. These distances represent the similarities between nuclei pairs in the graph. In our work we use the implemented algorithm in [31] which is introduced by [32]. In the paper authors minimize the Hamiltonian path and suggest a fast algorithm with time complexity  $O(n^4)$ .

- **Hierarchical Clustering Reordered by Gruvaeus and Wainer Algorithm (GW)**

The method reorders the objects with an additional criterion after performing hierarchical clustering. The proposed algorithm aims to find a unique optimal order of a binary hierarchical clustering tree by testing the arrangement of the leaf nodes so that, at each level the objects on the left and right edges of each cluster are adjacent to the nearest object outside the cluster; in this way, they are the most similar ones to each other. At this points our nuclei are the leaves of the hierarchical clustering tree and we aim to find an order where the most similar nuclei are side by side. In [31], package `gclus` [36] implementation is used for this ordering algorithm.

- **Traveling Salesperson Problem Solver (TSP)**

The traveling salesperson problem (TSP) is a famous optimization problem [37]. Ordering with TSP solver also corresponds to minimizing the Hamiltonian path length through a graph heuristically. In R seriation package

we use the algorithm which minimize the Hamilton path where the vertexes are our nuclei, and the edges of the graph are the distances between nucleus pairs.

- **Rank-two Ellipse Seriation by Chen**

In this ordering algorithm again the Hamiltonian path is the criteria in which the rank-two ellipse seriation method uses a minimal span loss function to calculate Hamiltonian path where the length of the Hamiltonian path is equal to the resulting value of the minimal span loss function [38].

- **ARSA**

ARSA is a heuristic simulated annealing algorithm for the ordering of objects which is included in the R seriation package [31]. A symmetric dissimilarity matrix in which the values in rows and columns only increase when moving away from the main diagonal is a perfect anti-Robinson matrix, and the number of violations in an anti-Robinson matrix are called as anti-Robinson events. The proposed algorithm aims to minimize anti-Robinson events as a loss function. In the [31], they use the code developed by [39] for this ordering method.

# Chapter 6

## Experiments

In this section we present and discuss our experimental segmentation and ordering results performed on the Hacettepe and Herlev data sets.

As described in Chapter 3 in detail, we first segment Pap smear images at 20x magnification. After we obtain segmented regions from 20x magnification Pap smear images, we eliminate non-nucleus regions by using four different features extracted from these segmented regions. Then, we segment the nucleus regions, which are obtained from 20x magnification, at 40x magnification to extract good quality features to use in the ordering procedure. Following segmentation, we order the segmented nucleus regions by using different features extracted from 40x magnification Pap smear images with different ordering algorithms.

Below we provide detailed experimental evaluations of our segmentation and ordering algorithms.

### 6.1 Evaluation of Segmentation

In the Hacettepe data set there are multiple cells; therefore, in order to evaluate the segmentation results of this data set, for each input image of Pap smear slide, we need to prepare a corresponding ground truth Pap smear image in which all

Table 6.1: The ZSI results of three Pap smear images for the ground truth compared to our segmentation.

40x Magnification Pap Smear Images	ZSIs
<i>Pap Smear Image 1</i>	0.857
<i>Pap Smear Image 2</i>	0.805
<i>Pap Smear Image 3</i>	0.780

nuclei boundaries are delineated manually. After preparation of ground truth images, in order to match and compare segmented nuclei in a Pap smear image with the corresponding ground truth image, we compute the Zijdenbos similarity index (ZSI) [40] which is basically the ratio of twice the common area between two regions  $A_1$  and  $A_2$  to the sum of individual areas. The Table 6.1 presents ZSI results of three Pap smear images from our Hacettepe data set. The segmentation results of 40x magnification *Pap Smear Image 1*, *Pap Smear Image 2* and *Pap Smear Image 3* are shown in Figure 6.2(a), 6.2(b) and 6.2(c) respectively. Since 0.7 specifies excellent agreement between the segments [40], it could be seen from the Table 6.1 that our proposed segmentation method is successful.

For the segmentation of nuclei in the Pap smear images, we proceed in two steps. First we segment Pap smear images at 20x magnification and select only the potential nucleus regions by eliminating non-nucleus regions with respect to thresholds values of four different features. Figure 6.1(a), 6.1(b), 6.1(c) show segmentation results for three different Pap smear images from the Hacettepe data set. In the second step, we segment the selected nucleus regions at corresponding 40x magnification Pap smear images as explained in *Chapter 3* in detail. Figure 6.2(a), 6.2(b), 6.2(c) show the final segmentation results of corresponding 40x magnification Pap smear images.

In our segmentation step, it is very important to obtain all the nuclei in the Pap smear images as accurately segmented nucleus regions. Once we have the true nucleus regions, our method is able to segment the most of those regions at 40x magnification; otherwise, once we miss a region at 20x magnification, we cannot obtain this region from the corresponding 40x magnification Pap smear image. Even though we obtain the most of nucleus regions successfully at 20x

magnification Pap smear images, we may miss or/and inaccurately segment some nuclei due to the some problems; especially the ones related to insufficient contrast. Below we specify these main problems.

- We could have insufficient contrast between nucleus and cytoplasm because of the nature of Pap smear images. The main reason of this insufficient contrast is the staining process of Pap smear images. As explained previously, after the staining procedure we come across inhomogeneity in a single slide. In such a case, usually the dye that is used to color the cells in Pap smear images spreads over a wider area in the cell and cause to smooth the contrast.
- Focus is one of the most important problems of our segmentation step that can be observed from the presented results. In a Pap smear image, there are thousands of cells and they overlap each other. To examine these cells which are overlapped by other cells, cyto-technicians use different focus levels. In order to make focus problem clear, in Figure 6.3, we present the same Pap smear image at three different focus levels. As it could be seen from the figure, at each focus level of Pap smear images we observe new cells form different layers of Pap smear slide.

In our 20x magnification Pap smear images, we have cells which are just dark regions, just like the cells in the middle of Figure 6.3(a). Even though, they are slightly darker than cytoplasm, they do not have shape of an ordinary nucleus. To understand what really these regions are, we need to switch to a different focus level. Our segmentation results also are affected by low contrast which is originated from this focus issue. In case of insufficient contrast, these regions are missed or segmented in such a way that they are merged with other darker regions.

Moreover, the focus level may not be the same for some nuclei at 20x and 40x magnification; so that, a nucleus, which could be seen clearly at 20x magnification, is invisible at 40x magnification, or vice versa. Since we use a single focus setting in the experiments, some of these problems cannot be avoided.

- Illumination is another problem which is related to microscope settings and the lighting condition of the room where the Pap smear images are captured. As an example, the illumination difference between Pap smears images could be seen in Figure 6.2(a) and Figure 6.2(b). In the elimination step of non-nucleus regions, we use the same threshold value of mean intensity feature for all Pap smear images in the data set. Therefore, after elimination, step we have some extra cytoplasm or background regions selected as nucleus regions due to this fact (see Figure 6.1 ). Mostly, except intensity, these regions are similar to a nucleus region in terms of size and circularity; therefore these regions should be eliminated. However, in such a case, because of the threshold value and their intensity values, they are considered as nucleus regions.
- The other factor which affects segmentation results is manually prepared 20x and 40x magnification Pap smear image pairs. Although we reduce the registration error by applying template matching, we still have this problem slightly in some Pap smear images (see Figure 6.1(b) and Figure 6.2(b)).

Among the mentioned issues, focus problem can be prevented by capturing multiple images of the same the Pap smear slide area at different focus levels. For the rest of these issues, which are beyond our control, using technologically more advanced devices such as camera and microscope could help to avoid illumination and drift problem. With them, segmentation results will be improved.

When we directly segment the Pap smear images at 40x magnification, we obtain over segmented results in which especially nuclei of cells are fragmented into many pieces due to more detailed texture of 40x magnification Pap smear images. However, we need detailed texture as a feature to be used in ordering step. Our segmentation method avoids this problem by obtaining nucleus regions from 20x magnification and extracting the features from corresponding regions of 40x magnification Pap smear images. Also, since sizes of 40x magnification images are four times larger than 20x magnification images, it takes considerably more time segmenting the Pap smear images at 40x magnification. Thus, our approach, segments nucleus regions accurately and extract good quality features

in computationally efficient way. Moreover, as a non-parametric, unsupervised and robust algorithm, the proposed segmentation method ensures the necessary conditions of an automatic screening system.

Since we have a single magnification for the Herlev data set we directly use the Herlev segmentation results from [21].

## 6.2 Evaluation of Ordering

In *Chapter 4* we explain our five different mapping functions for distance matrixes. After experimenting with these mapping functions, we observe that there is no significant difference between the simplest method (*Method 1*) and the other mapping functions in terms of performance. Therefore, in order to keep the process simple, we decide to use the *Method 1* which only shifts the distance matrix after z-score normalization.

Below, we present our ordering results for the segmented nucleus regions. Since we have the ground truth values of nuclei in the Herlev data set, we first show our ordering results for the Herlev data set.

### 6.2.1 Ordering Results of the Herlev Data Set

To measure the success of ordering algorithms statistically for the Herlev data set we use kappa coefficients as our evaluation criteria. Kappa coefficients, also known as Cohen's kappa coefficients, are statistical measures which quantify the correlation between categorical variables [41]. Since kappa coefficients take into account the agreement occurring by chance, it is considered to be a more robust measure compared to simple agreement calculation.

Suppose that each object in a set of  $N$  objects is assigned to one of  $g$  categories by two raters. Then, we get a confusion matrix  $n$  where  $n_{ij}^{\text{th}}$  element is the number of observations which are labeled as category  $i$  by the first rater and as category

$j$  by the second rater. Kappa coefficients are calculated as follows

$$k = \frac{P_o - P_e}{1 - P_e} \quad (6.1)$$

where  $P_o$  and  $P_e$  correspond to proportion of the observed and the expected agreement by the raters. Here  $k = 1$  and  $k = 0$  stand for complete agreement and no agreement among the raters, respectively.

### Weighted Kappa coefficients

Kappa coefficients only consider the matches between the observed agreement and the expected agreement on the main diagonal of confusion matrixes. Weighted kappa lets us consider off diagonal elements as well by including an additional weight matrix. Weighted kappa coefficients are calculated as

$$k_w = \frac{P_{o(w)} - P_{e(w)}}{1 - P_{e(w)}} = \frac{\sum w_{ij}P_{ij} - \sum w_{ij}P_iP_j}{1 - \sum w_{ij}P_iP_j}. \quad (6.2)$$

We use the following weight matrix as presented in [21]:

$$\begin{bmatrix} 1 & 0.5 & 0 & 0.25 & 0.25 & 0 & 0 \\ 0.5 & 1 & 0 & 0.25 & 0.25 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0 & 1 & 0.5 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0 & 0.5 & 1 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.25 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0 & 0.25 & 0.5 & 0.5 & 1 \end{bmatrix}. \quad (6.3)$$

To perform the ordering algorithms on the Herlev data set, we randomly select 25 cells from each class of the Herlev data set, except the columnar class. Since in the Hacettepe data set columnar cells are rarely encountered, we drop and do not include the columnar cells in our ordering experiments. Below we evaluate the ordering results of these 150 cells from the Herlev data set.

In Table 6.2, we show our single and combined features which are used in the ordering algorithms. The features between F1-F8 correspond to the features

defined by us. The features between F9-F15 are a subset of the features used in [4] for characterizing cervical cells. Rest of the features, between F16-F26, are different combinations of the given single features. In the following we explain the details of the combined features.

- $k$  and  $k_w$  performances of the circularity, the contrast and the homogeneity features are worse than 0.2 for all of the ordering algorithms. Therefore, in order to improve the ordering algorithms performances, we construct a new subset of our features called F17, which does not include circularity, contrast and homogeneity features.
- F16 is similar to F17; but this time, it does not include mean intensity values of a and b channels of the Lab color space. In other words, we only have the L channel of the Lab color space as a mean intensity feature.
- F18, F19 and F20 respectively include the circularity, the contrast and the homogeneity features in addition to F17 in order to show how they affect the ordering algorithms performances.
- F21 includes only the features between F9-F15 which are a subset of the features used in [4].
- F23 is the combination of F17 and F21 in which there are features mostly with the better performance than 0.2 in terms of  $k$  and  $k_w$ .
- F22 is similar to F23, but this time, it does not include mean intensity values of a and b channels of the Lab color space. In other words, we only have the L channel of the Lab color space as a mean intensity feature.
- F24, F25 and F26 respectively include the circularity, the contrast and the homogeneity features in addition to F23.

The Tables 6.3, 6.4, 6.5, 6.6, 6.7 and 6.8 respectively show the performances of the ordering algorithms HC, OLO, GW, TSP, Chen and ARSA. Table 6.9 shows the best performance analysis of each of these ordering algorithm with the corresponding feature set(s), and Figures 6.4, 6.5, 6.6, 6.7, 6.8 and 6.9 illustrate

these best performances on the input Herlev data set. In the experimental results higher values of  $k$  and  $k_w$  indicate better ordering performances.

As the results in the Table 6.9 show, we get the best performance with the features defined by us. According to the ordering algorithm performances presented in the tables, the L channel mean intensity of the Lab color space, size and texture features of a nucleus region are essential features to determine its abnormality degree as expected (F16). Using a and b channel mean intensity values of the Lab color space improves the ordering performances (F17). The most interesting thing could be observed from the results is that; circularity feature do not show good performance alone (F3). However when it is used with the essential features, which are L channel mean intensity of the Lab color space, size and texture, the performances of the ordering algorithms increase significantly (F18). Since circularity represents the deformation in the boundaries of the nucleus regions, this case is also reasonable.

Among the ordering algorithms, the worst performance belongs to HC (Hierarchical Clustering) algorithm and the best performances belong to GW and OLO algorithms for the most of the feature combinations. Basically, both GW and OLO are extended versions of HC algorithm. OLO is the hierarchical clustering reordered by optimal leaf ordering and GW is the hierarchical clustering reordered by Gruvaeus and Wainer algorithm as explained in Chapter 5.

These results indicate that we could get an ordered list of nuclei, without cytoplasm features by using only the features extracted from nuclei. Moreover, Figures 6.4, 6.5, 6.6, 6.7, 6.8 and 6.9 show that in the results of the ordering algorithms, we have naturally formed two groups of normal and abnormal cells, where normal cells are located in the one end of the ordered list and the abnormal cells are located in the other end of the ordered list.

In [21], the best performance values obtained by using only nucleus features are  $k = 0.055$  and  $k_w = 0.140$ . This shows that our features and ordering algorithms improves the ordering results. Moreover, we could obtain more improved ordering results by using cytoplasm features together with our nucleus features.

Table 6.2: Features for the Herlev Data Set

Features	Explanations
F1	mean intensity
F2	size
F3	circularity
F4	contrast
F5	homogeneity
F6	lbp
F7	mean intensity of a-channel
F8	mean intensity of b-channel
F9	nucleus elongation
F10	nucleus roundness
F11	nucleus perimeter
F12	nucleus longest diameter
F13	nucleus shortest diameter
F14	nucleus maxima
F15	nucleus minima
F16	$F1 + F2 + F6$
F17	$F1 + F2 + F6 + F7 + F8$
F18	$F1 + F2 + F3 + F6 + F7 + F8$
F19	$F1 + F2 + F4 + F6 + F7 + F8$
F20	$F1 + F2 + F5 + F6 + F7 + F8$
F21	$F9 + F10 + F11 + F12 + F13 + F14 + F15$
F22	$F1 + F2 + F6 + F9 + F10 + F11 + F12 + F13 + F14 + F15$
F23	$F1 + F2 + F6 + F7 + F8 + F9 + F10 + F11 + F12 + F13 + F14 + F15$
F24	$F1 + F2 + F3 + F6 + F7 + F8 + F9 + F10 + F11 + F12 + F13 + F14 + F15$
F25	$F1 + F2 + F4 + F6 + F7 + F8 + F9 + F10 + F11 + F12 + F13 + F14 + F15$
F26	$F1 + F2 + F5 + F6 + F7 + F8 + F9 + F10 + F11 + F12 + F13 + F14 + F15$

Table 6.3: HC Ordering Performance

Features	$k$	$k_w$
F1	0.072	0.266
F2	0.176	0.313
F3	0.024	0.005
F4	-0.016	0.007
F5	-0.016	0.007
<b>F6</b>	<b>0.288</b>	<b>0.411</b>
F7	0.072	0.266
F8	0.072	0.266
F9	0.072	0.266
F10	0.072	0.266
F11	0.072	0.266
F12	0.072	0.266
F13	0.072	0.266
F14	0.056	0.253
F15	0.072	0.264
F16	0.144	0.253
F17	0.072	0.133
F18	-0.008	0.204
F19	0.080	0.105
F20	0.080	0.105
F21	0.072	0.266
F22	0.016	0.111
F23	0.072	0.264
F24	0.152	0.296
F25	0.008	0.040
F26	0.008	0.040

Table 6.4: OLO Ordering Performance

Features	$k$	$k_w$
F1	0.232	0.378
F2	0.232	0.378
F3	0.104	0.187
F4	0.032	0.024
F5	0.032	0.024
F6	0.240	0.384
F7	0.232	0.378
F8	0.232	0.378
F9	0.232	0.378
F10	0.232	0.378
F11	0.232	0.378
F12	0.232	0.378
F13	0.232	0.378
F14	0.232	0.378
F15	0.232	0.378
F16	0.248	0.389
F17	0.248	0.386
<b>F18</b>	<b>0.296</b>	<b>0.425</b>
F19	0.128	0.269
F20	0.128	0.269
F21	0.232	0.378
F22	0.240	0.378
F23	0.256	0.386
F24	0.192	0.335
F25	0.232	0.378
F26	0.232	0.378

Table 6.5: GW Ordering Performance

Features	$k$	$k_w$
F1	0.272	0.408
F2	0.288	0.408
F3	0.104	0.187
F4	0.032	0.026
F5	0.032	0.026
F6	0.240	0.381
F7	0.272	0.408
F8	0.272	0.408
F9	0.272	0.408
F10	0.272	0.408
F11	0.272	0.408
F12	0.272	0.408
F13	0.272	0.408
F14	0.240	0.381
F15	0.240	0.381
F16	0.216	0.335
F17	0.224	0.370
<b>F18</b>	<b>0.328</b>	<b>0.425</b>
F19	0.184	0.307
F20	0.184	0.307
F21	0.272	0.408
F22	0.232	0.378
F23	0.240	0.381
F24	0.208	0.351
F25	0.184	0.343
F26	0.184	0.343

Table 6.6: TSP Ordering Performance

Features	$k$	$k_w$
F1	0.280	0.397
F2	0.232	0.378
F3	0.104	0.187
F4	0.040	0.032
F5	0.048	0.037
F6	0.248	0.389
F7	0.248	0.375
F8	0.280	0.400
F9	0.208	0.335
F10	0.288	0.403
F11	0.240	0.378
<b>F12</b>	<b>0.320</b>	<b>0.425</b>
F13	0.304	0.422
F14	0.056	0.231
F15	0.248	0.373
F16	0.288	0.403
F17	0.288	0.414
F18	0.000	0.149
F19	0.208	0.321
F20	0.160	0.280
F21	0.312	0.425
F22	0.216	0.329
F23	0.272	0.395
F24	0.264	0.381
F25	0.224	0.367
F26	0.232	0.367

Table 6.7: Chen Ordering Performance

Features	$k$	$k_w$
F1	0.256	0.384
<b>F2</b>	<b>0.256</b>	<b>0.386</b>
F3	0.104	0.187
F4	0.024	0.015
F5	0.024	0.015
F6	0.248	0.378
F7	0.256	0.384
F8	0.256	0.384
F9	0.256	0.384
F10	0.256	0.384
F11	0.256	0.384
F12	0.256	0.384
F13	0.256	0.384
F14	0.256	0.384
F15	0.256	0.384
<b>F16</b>	<b>0.256</b>	<b>0.386</b>
<b>F17</b>	<b>0.256</b>	<b>0.386</b>
F18	0.232	0.365
F19	0.192	0.340
F20	0.192	0.340
F21	0.256	0.384
F22	0.248	0.378
F23	0.248	0.378
F24	0.248	0.378
F25	0.240	0.375
F26	0.240	0.375

Table 6.8: ARSA Ordering Performance

Features	$k$	$k_w$
F1	0.248	0.378
F2	0.248	0.384
F3	0.104	0.187
F4	0.040	0.032
F5	0.024	0.015
F6	0.248	0.378
F7	0.248	0.378
F8	0.248	0.378
F9	0.248	0.378
F10	0.248	0.378
F11	0.248	0.378
F12	0.248	0.378
F13	0.248	0.378
F14	0.248	0.378
F15	0.248	0.378
F16	0.248	0.378
F17	0.248	0.378
<b>F18</b>	<b>0.264</b>	<b>0.395</b>
F19	0.208	0.356
F20	0.208	0.356
F21	0.248	0.378
F22	0.248	0.378
F23	0.248	0.378
F24	0.256	0.386
F25	0.248	0.378
F26	0.248	0.378

Table 6.9: Best Performance Analyses of the Ordering Algorithms

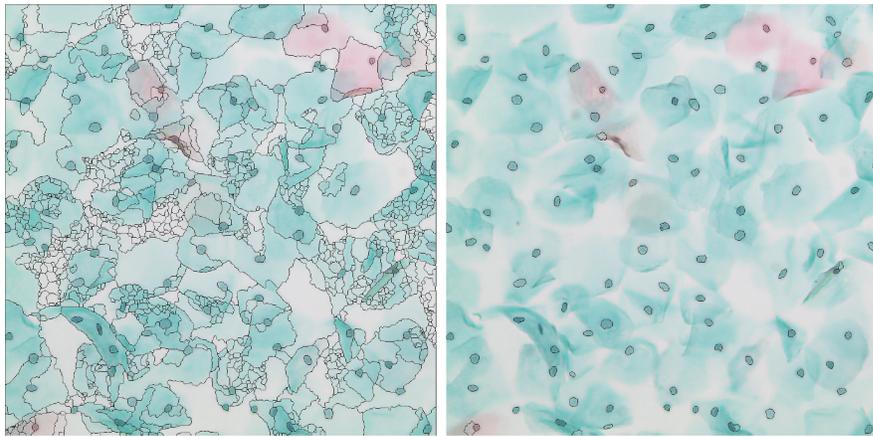
The Ordering Algorithm	The Best Performance Feature Set	$k$	$k_w$
HC	F6	0.288	0.411
OLO	F18	0.296	0.425
<b>GW</b>	<b>F18</b>	<b>0.328</b>	<b>0.425</b>
TSP	F12	0.320	0.414
Chen	F2, F16, F17	0.256	0.425
ARSA	F18	0.264	0.395

### 6.2.2 Ordering Results of the Hacettepe Data Set

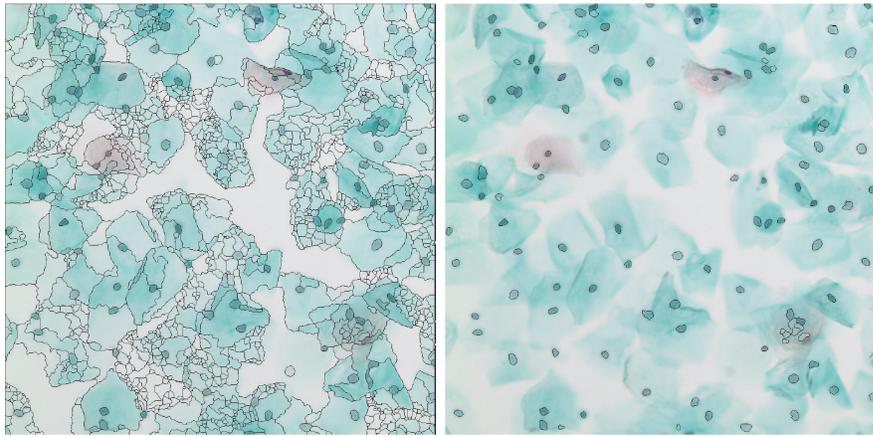
Figure 6.10 shows an example ordering result for a Pap smear image from the Hacettepe data set. This resulting image is obtained by applying OLO ordering algorithm with a new feature set that includes L channel mean intensity of the Lab color space, size, circularity and texture features (F1+F2+F3+F6). Since we do not have the ground truth abnormality degree values of the cells in the Hacettepe data set, we are not able to present numeric results. However, as it could be observed from the figure, the color bar specifies the order of the nucleus regions where blue color and its tones correspond to potential large abnormal nucleus regions and red color and its tones correspond to small nucleus regions or inflammations.

## 6.3 Implementation Settings and Computational Complexity

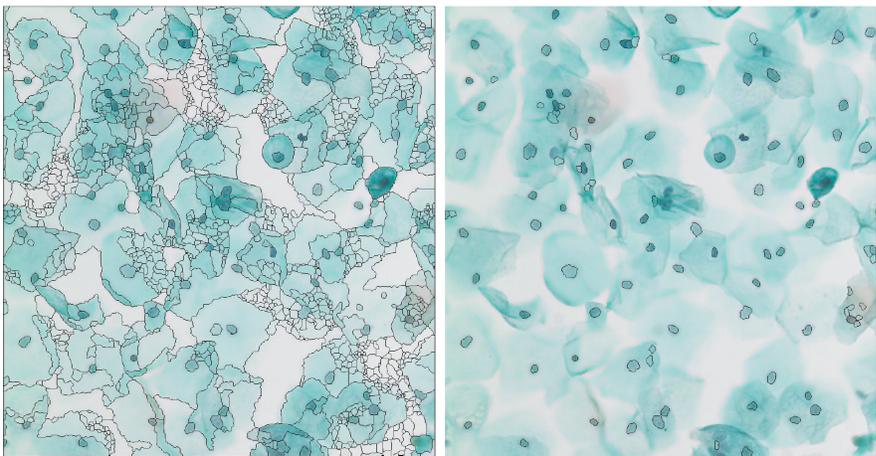
Template matching process of the segmentation step was implemented in *C++* by using the *OpenCV* library. Rest of the segmentation step was implemented in *Matlab*. The ordering step was implemented in *R Project* by applying the ordering algorithms that are provided by the *R Package seriation*. In a PC with a 2.30 GHz Intel Core i7 processor and 8 GB RAM, the overall process of segmenting nucleus regions at 20x magnification and ordering these segmented nucleus regions in 40x magnification takes 1280 seconds ( $\sim 21$  minutes) for an example Pap smear pair from the Hacettepe data set which includes 98 segmented nuclei. Time complexity can be improved by using optimized codes for the Matlab implementation part.



(a)

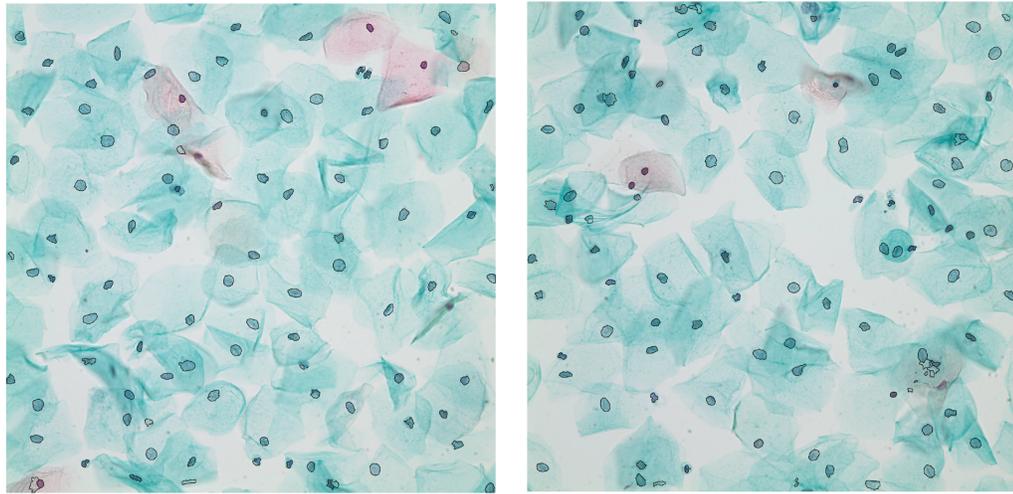


(b)



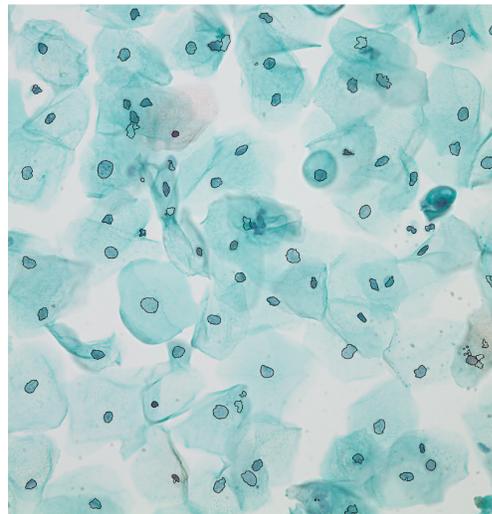
(c)

Figure 6.1: Segmentation results of three Pap smear images at 20x magnification. 1<sup>st</sup> column shows initial segmented results and 2<sup>nd</sup> row shows the selected nucleus regions.



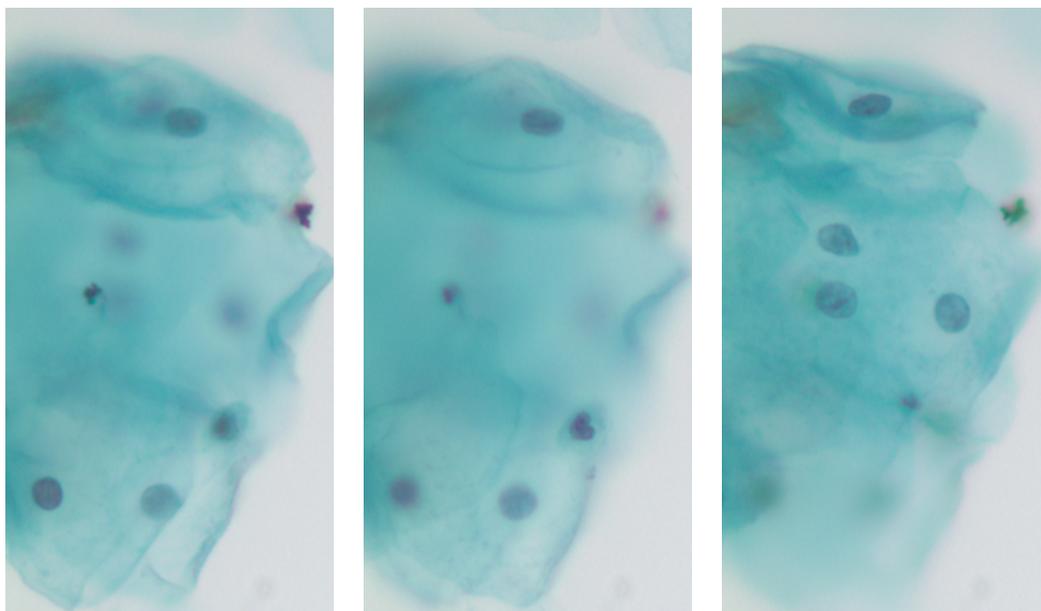
(a)

(b)



(c)

Figure 6.2: Final segmentation results at 40x magnification. The given images are corresponding pairwise images of 20x magnification Pap smear images shown in Figure 6.1.



(a)

(b)

(c)

Figure 6.3: Three Pap smear images that correspond to the same Pap smear slide area at three different focus settings.

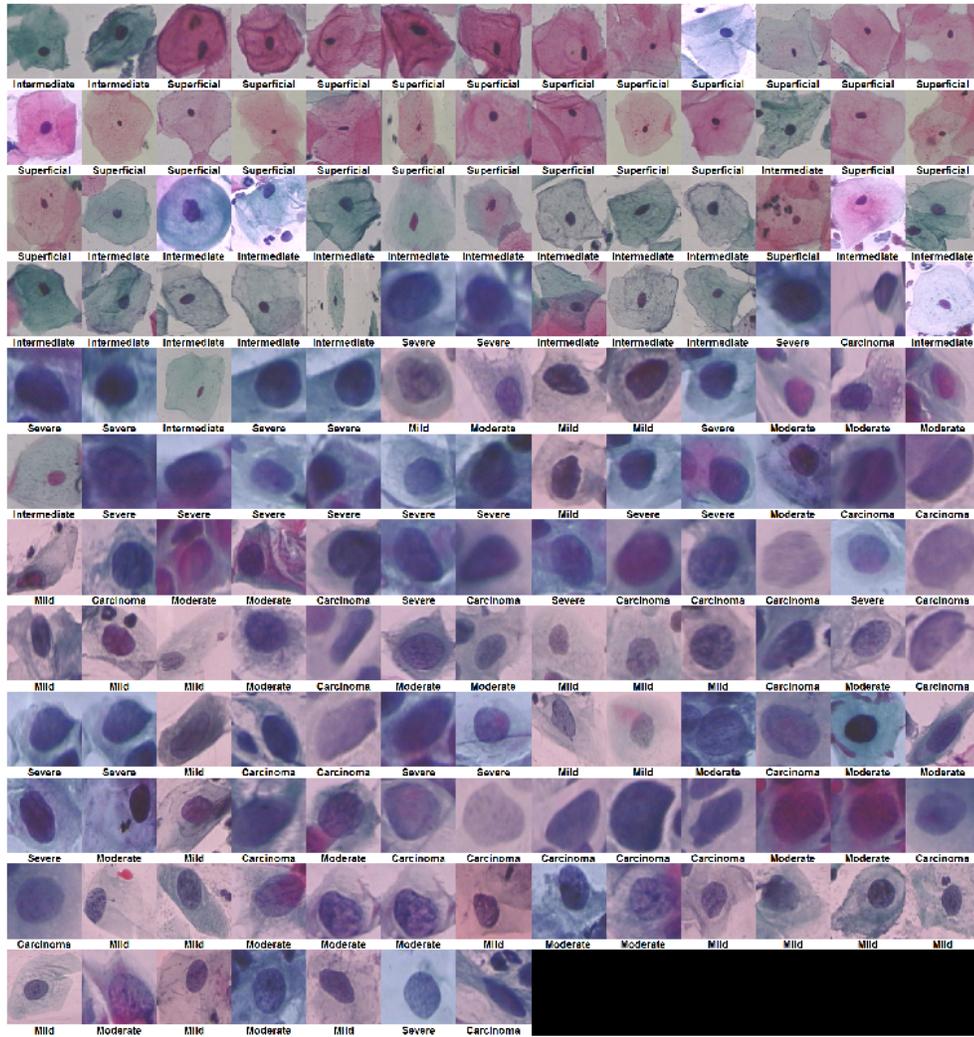


Figure 6.4: Best performance of the ordering algorithm HC:  $k = 0.288$ ,  $k_w = 0.411$ . The images are resized to the same width and height so the relative sizes of the cells are not proper.

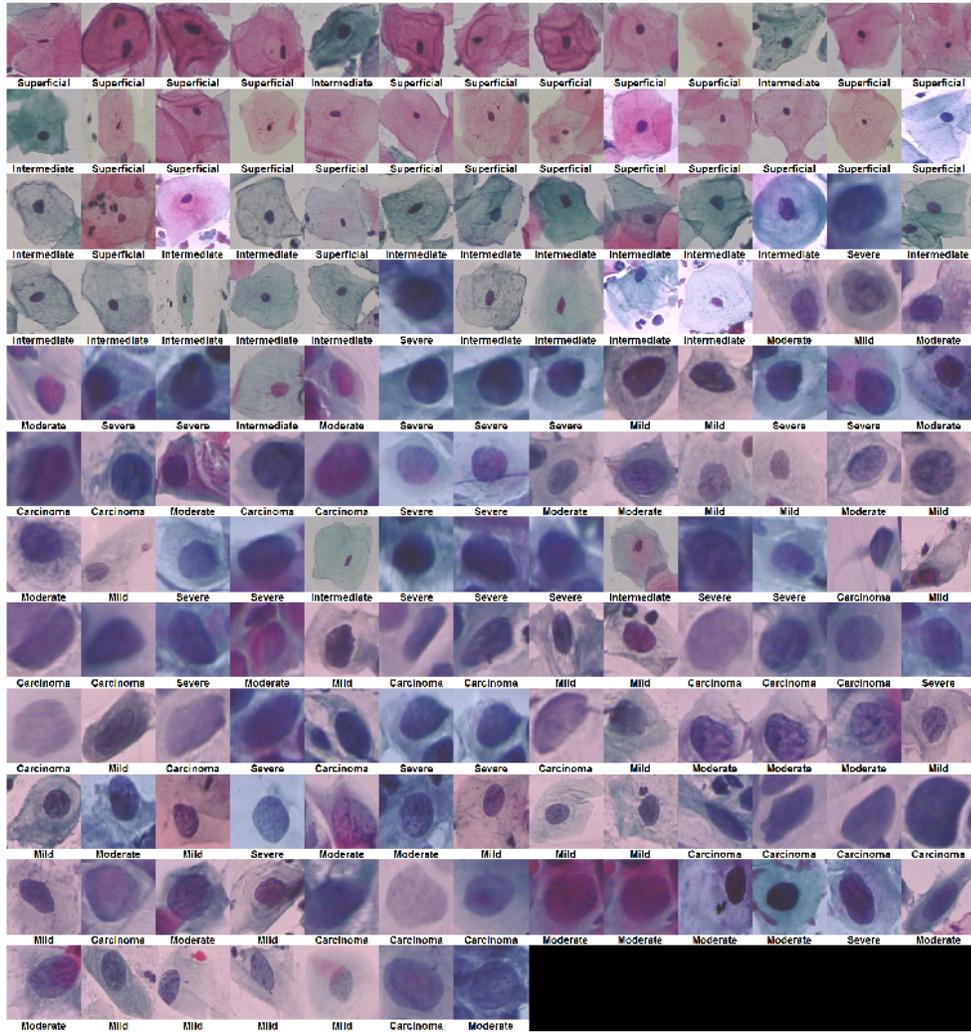


Figure 6.5: Best performance of the ordering algorithm OLO:  $k = 0.296$ ,  $k_w = 0.425$ . The images are resized to the same width and height so the relative sizes of the cells are not proper.

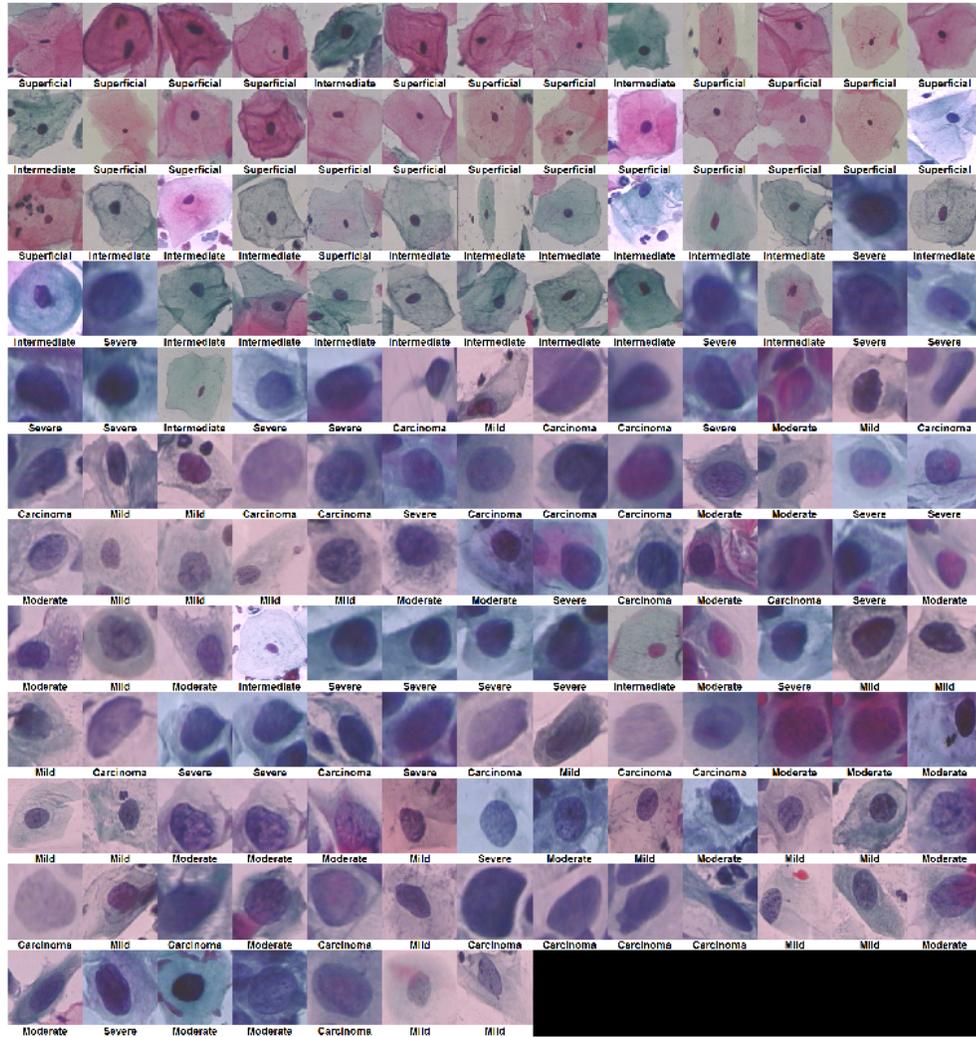


Figure 6.6: Best performance of the ordering algorithm GW:  $k = 0.328$ ,  $k_w = 0.425$ . The images are resized to the same width and height so the relative sizes of the cells are not proper.

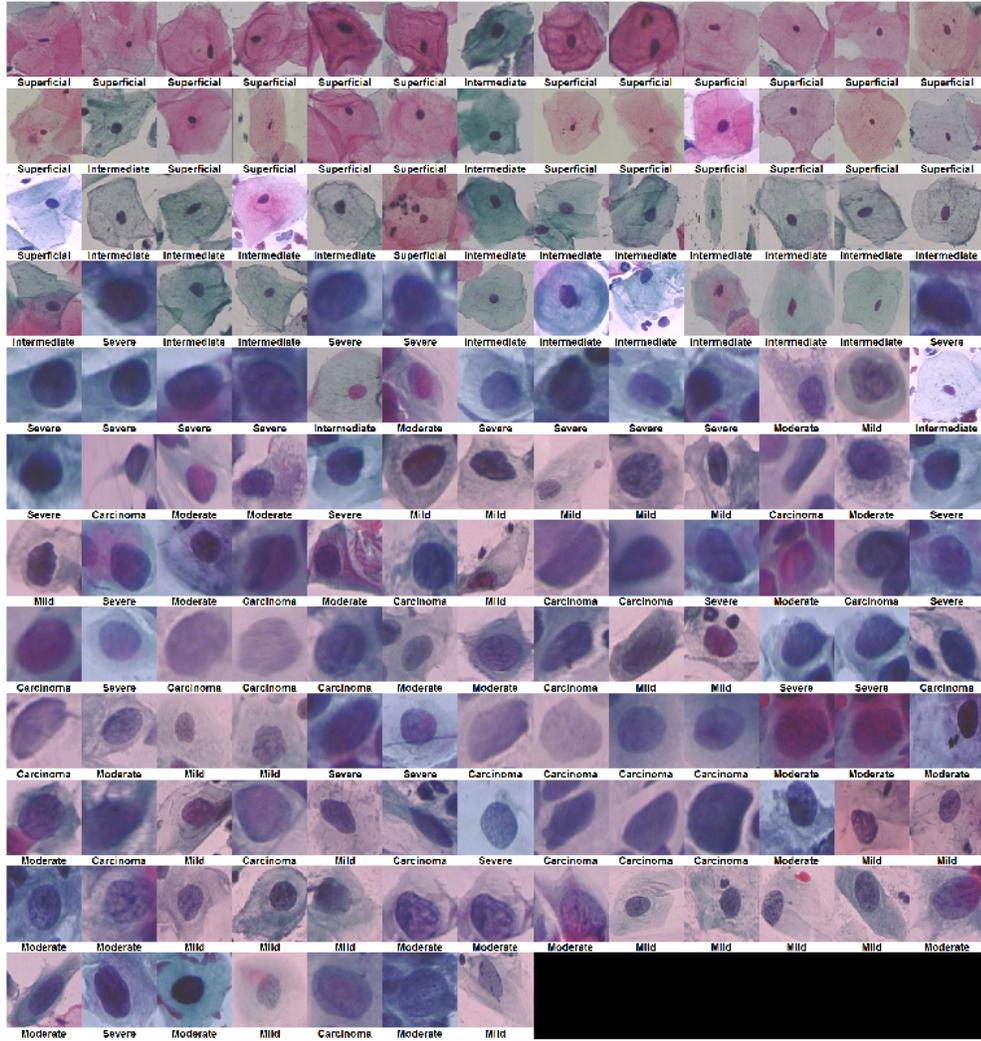


Figure 6.7: Best performance of the ordering algorithm TSP:  $k = 0.288$ ,  $k_w = 0.414$ . The images are resized to the same width and height so the relative sizes of the cells are not proper.

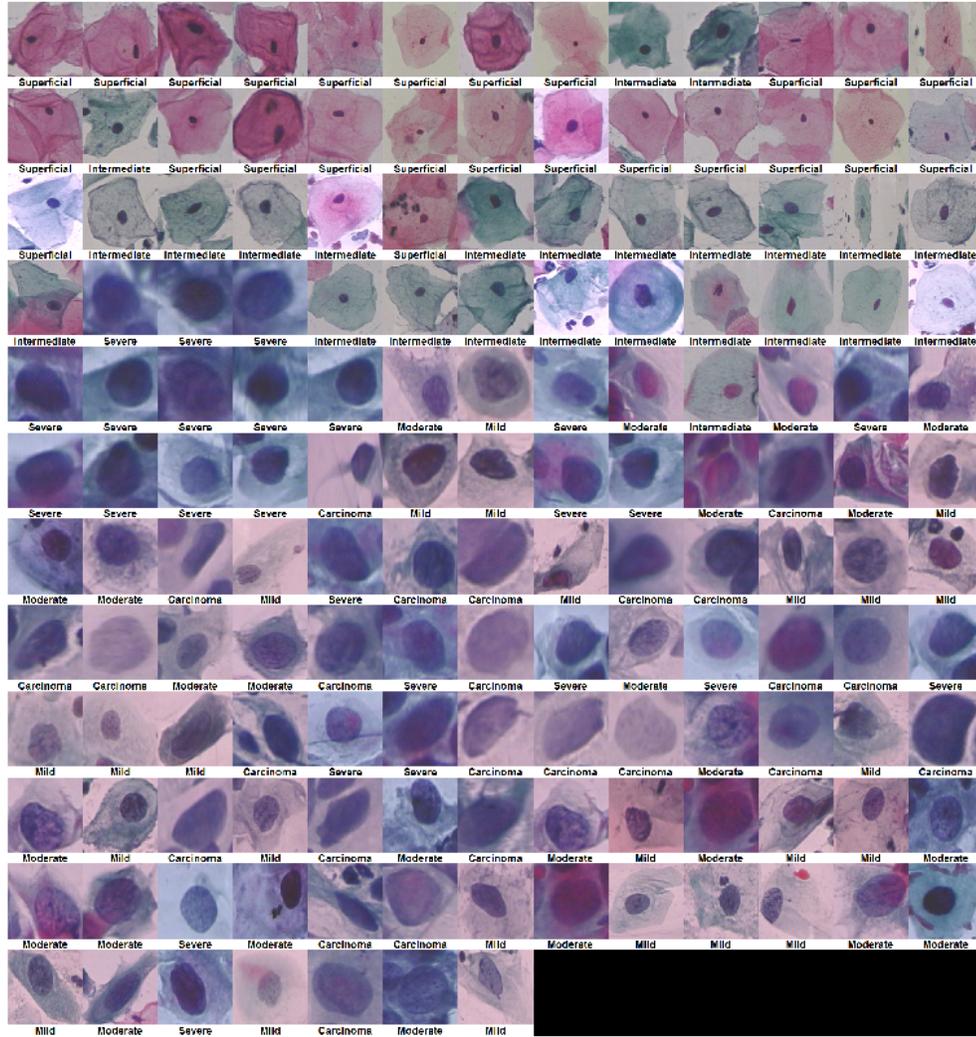


Figure 6.8: Best performance of the ordering algorithm Chen:  $k = 0.256$ ,  $k_w = 0.386$ . The images are resized to the same width and height so the relative sizes of the cells are not proper.

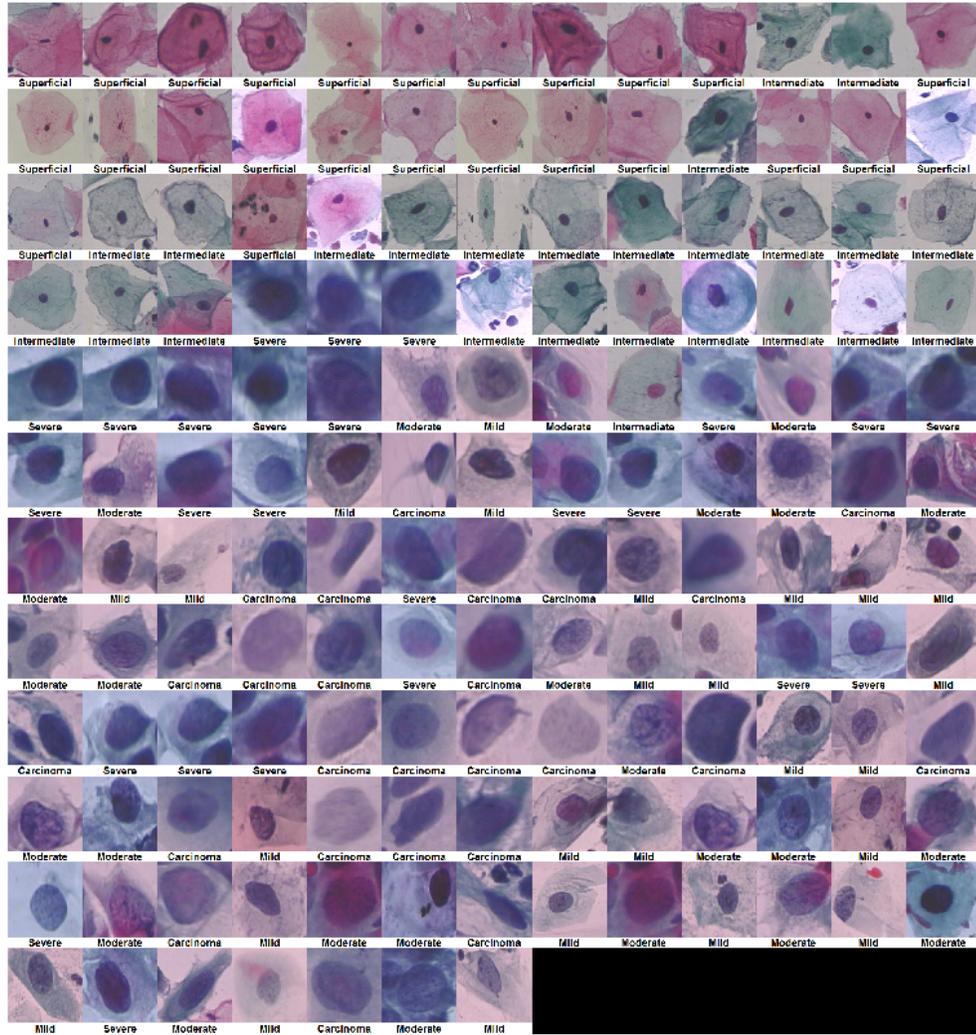


Figure 6.9: Best performance of the ordering algorithm ARSA:  $k = 0.264$ ,  $k_w = 0.395$ . The images are resized to the same width and height so the relative sizes of the cells are not proper.

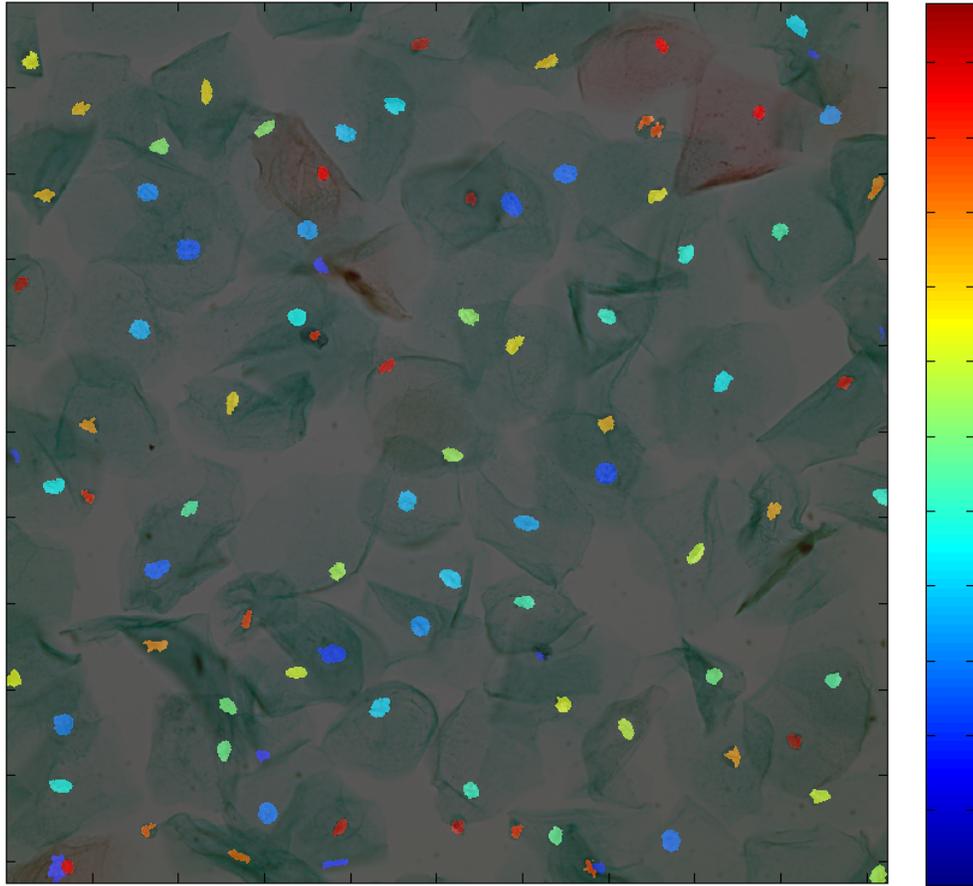


Figure 6.10: Ordering result for the segmented nucleus regions of Figure 6.2(a). The tones of colors represent the similarities between the nucleus regions.

# Chapter 7

## Conclusion

We presented a computer-assisted screening procedure which aims to automate the Pap smear screening process and provide an ordered nuclei list to help the cyto-experts. We only aimed to accurately segment nucleus regions by using a multi-scale hierarchical segmentation algorithm. To overcome contrast and detailed texture tradeoff, where contrast is an important property for segmentation and detailed texture is an important property for feature extraction, we first segmented the Pap smear images with better contrast at low level (20x) magnification, and we only chose the regions which are considered as nucleus regions. Then, we switched to a higher level (40x) magnification of the Pap smear images to extract good quality features for each obtained nucleus, in terms of morphological properties of the nucleus regions, i.e. size, color, shape and texture. Finally we compared different ordering algorithms using the features extracted from nucleus regions for the ranking of the segmented nucleus regions according to their abnormality degrees.

Considering the fact that it is even very difficult for an expert to differentiate the boundaries of cytoplasm for each nucleus, where grouped cells overlap and occlude each other in the Pap smear images, our ultimate goal is the ordering of the cervical cells by using only the features extracted from the segmented nucleus regions. Our experiments using two data sets showed that the proposed unsupervised segmentation and ordering methods could accurately segment and

order the segmented nucleus regions according to their abnormality degree in images having inconsistent staining, poor contrast, and unknown number of cells. Furthermore, the experiments also showed that the cervical cells could be sorted according to their abnormality degree by using only the nucleus features. The ordering algorithms produced a list where the normal cells are located in one end of the list and the abnormal cells are located in the other end of the list. In this way, the ordered list groups the cells into normal and abnormal classes.

# Bibliography

- [1] “World health organization statistical information system.” <http://www.who.int/mediacentre/factsheets/fs380/en/>, 2013.
- [2] G. N. Papanicolaou, “A new procedure for staining vaginal smears,” *Science (New York, N.Y.)*, 1942.
- [3] “Health report fiscal years 2005-2006,” *North Carolina Institute*, 2007.
- [4] E. Martin., “Pap-smear classification,” 2003. Master’s Thesis, Technical University of Denmark: Oersted-DTU, Automation.
- [5] S.-F. Yang-Mao, Y.-K. Chan, and Y.-P. Chu, “Edge enhancement nucleus and cytoplasm contour detector of cervical smear images,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, pp. 353–366, 2008.
- [6] M.-H. Tsai, Y.-K. Chan, Z.-Z. Lin, S.-F. Yang-Mao, and P.-C. Huang, “Nucleus and cytoplasm contour detector of cervical smear image,” *Pattern Recognition Letters*, pp. 1441–1453, 2008.
- [7] P. Bamford and B. Lovell, “Unsupervised cell nucleus segmentation with active contours,” *Signal Processing*, vol. 71, no. 2, pp. 203 – 213, 1998.
- [8] I. Dagher and K. E. Tom, “Waterballoons: A hybrid watershed balloon snake segmentation,” *Image and Vision Computing*, vol. 26, no. 7, pp. 905 – 912, 2008.
- [9] P.-W. Huang and Y.-H. Lai, “Effective segmentation and classification for {HCC} biopsy images,” *Pattern Recognition*, vol. 43, no. 4, pp. 1550 – 1563, 2010.

- [10] N. M. Harandi, S. Sadri, N. A. Moghaddam, and R. Amirfattahi, “An automated method for segmentation of epithelial cervical cells in images of thinprep,” *J. Medical Systems*, pp. 1043–1058, 2010.
- [11] K. Li, Z. Lu, W. Liu, and J. Yin, “Cytoplasm and nucleus segmentation in cervical smear images using radiating {GVF} snake,” *Pattern Recognition*, vol. 45, no. 4, pp. 1255 – 1264, 2012.
- [12] M. E. Plissiti, C. Nikou, and A. Charchanti, “Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 2, pp. 233–241, 2011.
- [13] M. E. Plissiti, C. Nikou, and A. Charchanti, “Combining shape, texture and intensity features for cell nuclei extraction in pap smear images,” *Pattern Recognition Letters*, vol. 32, no. 6, pp. 838–853, 2011.
- [14] S. Ali and A. Madabhushi, “Active contour for overlap resolution using watershed based initialization (acorew): Applications to histopathology,” in *Biomedical Imaging: 2011 IEEE International Symposium on From Nano to Macro*, pp. 614–617, March 2011.
- [15] H.-S. Wu, J. Barba, and J. Gil, “A parametric fitting algorithm for segmentation of cell images,” *IEEE Transactions on Biomedical Engineering*, vol. 45, pp. 400–407, March 1998.
- [16] R. Walker, P. Jackway, B. Lovell, and I. D. Longstaff, “Classification of cervical cell nuclei using morphological segmentation and textural feature extraction,” in *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, 1994.*, pp. 297–301, Nov 1994.
- [17] S. Shah, “Automatic cell image segmentation using a shape-classification model,” in *Machine Vision and its Applications*, pp. 428–432, 2007.
- [18] T. Chankong, N. Theera-Umpon, and S. Auephanwiriyaikul, “Automatic cervical cell segmentation and classification in pap smears,” *Computer Methods and Programs in Biomedicine*, vol. 113, no. 2, pp. 539 – 556, 2014.

- [19] Z. Lu, G. Carneiro, and A. P. Bradley, “Automated nucleus and cytoplasm segmentation of overlapping cervical cells,” in *Medical Image Computing and Computer-Assisted Intervention*, vol. 8149 of *Lecture Notes in Computer Science*, pp. 452–460, Springer, 2013.
- [20] Y.-Y. Wang, Y.-N. Sun, C.-C. K. Lin, and M.-S. Ju, “Nerve cell segmentation via multi-scale gradient watershed hierarchies,” in *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE*, vol. Supplement, pp. 6698–6701, Aug 2006.
- [21] A. Gençtav, S. Aksoy, and S. Önder, “Unsupervised segmentation and classification of cervical cell images,” *Pattern Recognition*, vol. 45, no. 12, pp. 4151–4168, 2012.
- [22] N. Theera-Umpon, “White blood cell segmentation and classification in microscopic bone marrow images,” in *FSKD (2)* (L. Wang and Y. Jin, eds.), vol. 3614 of *Lecture Notes in Computer Science*, pp. 787–796, Springer, 2005.
- [23] Y.-Y. Chou and L. G. Shapiro, “A hierarchical multiple classifier learning algorithm,” *Pattern Anal. Appl.*, vol. 6, no. 2, pp. 150–168, 2003.
- [24] J. Zhang and Y. Liu, “Cervical cancer detection using svm based feature screening,” in *Proc. Seventh Intl Conf. Medical Image Computing and Computer Aided Intervention*, pp. 873–880, 2004.
- [25] Y. Marinakis, G. Dounias, and J. Jantzen, “Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification,” *Computers in Biology and Medicine*, vol. 39, no. 1, pp. 69 – 78, 2009.
- [26] H. G. Akcay and S. Aksoy, “Automatic detection of geospatial objects using multiple hierarchical segmentations,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 7, pp. 2097–2111, 2008.
- [27] D.-C. He and L. Wang, “Texture features based on texture spectrum,” *Pattern Recognition*, vol. 24, no. 5, pp. 391 – 399, 1991.

- [28] L. Wang and D.-C. He, “Texture classification using texture spectrum,” *Pattern Recognition*, vol. 23, no. 8, pp. 905 – 910, 1990.
- [29] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971–987, Jul 2002.
- [30] T. Ojala, M. Pietikainen, and T. Maenpaa, “A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification,” in *Advances in Pattern Recognition, ICAPR 2001*, vol. 2013 of *Lecture Notes in Computer Science*, pp. 399–408, 2001.
- [31] M. Hahsler, K. Hornik, and C. Buchta, “Getting things in order: An introduction to the r package seriation,” Research Report Series / Department of Statistics and Mathematics 58, Department of Statistics and Mathematics, WU Vienna University of Economics and Business, Vienna, 2007.
- [32] Z. Bar-Joseph, D. K. Gifford, and T. Jaakkola, “Fast optimal leaf ordering for hierarchical clustering.,” in *Bioinformatics-Oxford*, pp. 22–29, 2001.
- [33] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns,” *Proceedings of the National Academy of Sciences*, vol. 95, pp. 14863–14868, Dec. 1998.
- [34] H. C. Causton, B. Ren, S. S. Koh, C. T. Harbison, E. Kanin, E. G. Jennings, T. I. Lee, H. L. True, E. S. Lander, and R. A. Young, “Remodeling of yeast genome expression in response to environmental changes,” *Molecular Biology of the Cell*, vol. 12, pp. 323–337, Feb 2001.
- [35] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher, “Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization,” *Mol. Biol. of the Cell*, vol. 9, pp. 3273–3297, 1998.
- [36] C. Hurley, “gclus: Clustering graphics. r package version 1.2,” 2007.

- [37] G. Gutin, A. Punnen, A. Barvinok, E. K. Gimadi, and A. I. Serdyukov, “The traveling salesman problem and its variations,” 2002.
- [38] C. CH, “Generalized association plots: information visualization via iteratively generated correlation matrices,” *Statistica Sinica*, vol. 12, pp. 7–29, 2002.
- [39] M. Brusco, H.-F. Khn, and S. Stahl, “Heuristic Implementation of Dynamic Programming for Matrix Permutation Problems in Combinatorial Data Analysis,” *Psychometrika*, vol. 73, pp. 503–522, September 2008.
- [40] 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*, vol. 13, pp. 716–724, Dec 1994.
- [41] J. Cohen, “A Coefficient of Agreement for Nominal Scales,” *Educational and Psychological Measurement*, vol. 20, no. 1, p. 37, 1960.