ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

MSc THESIS

Çağlar CENGİZLER

A FLUID DYNAMICS BASED IMAGE SEGMENTATION APPROACH AND PAP-SMEAR IMAGE DATA CLASSIFICATION

DEPARTMENT OF COMPUTER ENGINEERING

ADANA, 2013

ABSTRACT

MSc THESIS

A FLUID DYNAMICS BASED IMAGE SEGMENTATION APPROACH AND PAP-SMEAR IMAGE DATA CLASSIFICATION

Çağlar CENGİZLER
ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF COMPUTER ENGINEERING

Supervisor : Assoc. Prof. Dr. Mutlu Avcı
Co-supervisor : Assoc. Prof. Dr. Mustafa Güven
Year: 2013 Pages 77
Jury : Assoc. Prof. Dr. Mutlu Avcı
: Asst. Prof. Dr. Mustafa Oral
: Asst. Prof. Dr. Cabbar Veysel Baysal

Today, cancer is one of the most important health issues of humanity. Diagnosis of cancer is based on examination of tissue samples, taken from patients by pathologists. One of the most common types of cancer is cervical cancer and it' s diagnosis is made by Pap-smear test which is based on visual examinations of biopsy samples under microscope for detection of anomalies.

In order to reduce human error and accelerate the Pap-smear test, computer based decision and detection systems are required.

In this study, development of a computer aided analyses system is aimed which is capable of detect anomalies on Pap-smear samples. Morphological features of cells on the samples, such as growth of cell nucleus or deformation on shape of cytoplasm area may provide crucial information about existence of cancer to such a computer aided system. However, in most cases of pap-smear image data, cytoplasm areas are overlapped and there are several types of artifacts on the samples which makes it difficult to extract features from cells for detection of abnormalities. Therefore a new segmentation method is developed and used in this study in addition to conventional morphological methods.

Proposed method is based on fluid dynamics. Modeled fluid during the segmentation process is capable of penetrate entire cytoplasm areas even where high degree of cell overlapping occurs. In this way, extracted features form cell are used for classification after segmentation process. Besides, some of the machine learning algorithms are examined and compared in this study for classification of Pap-smear samples for development of a analyses system which is capable of recognize abnormalities on sample cells.

Keywords: Pap-smear, *classification*, diagnosis, fluid dynamics

ÖZ

MASTER TEZİ

PAP-SMEAR ÖRNEKLERINDE AKIŞKAN DINAMIĞINE DAYALI BIR SEGMENTASYON YAKLAŞIMI VE GÖRÜNTÜ VERISININ SINIFLANDIRILMASI

Çağlar CENGİZLER

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Danışman : Doç Dr. Mutlu AVCI
2. Danışman : Doç. Dr. Mustafa Güven
Yıl: 2013 Sayfa 77
Jüri : Doç. Dr. Mutlu Avcı
Yrd. Doç. Dr. Mustafa Oral
Yrd. Doç. Dr. Cabbar Veysel Baysan

Kanser, günümüzde insanlığın en önemli sağlık sorunlarından birisidir. Kanserin kesin tanısı hastadan alınan doku örneklerinin patologlar tarafından incelenmesi ile konulmaktadır. Yaygın olarak görülen kanser türlerinden birisi de, serviks kanseridir ve teşhisi, Pap-smear testleri aracılığı ile yapılmaktadır. Pap-smear testi, hastadan alınan hücre örneklerinin mikroskop altında patologlar tarafından gözle incelenerek çeşitli anomalilerin aranmasına dayanır. Pap-smear örneklerinde, insan hatasının azaltılabilmesi ve test sürecinin hızlandırılabilmesi amacıyla bilgisayar destekli karar verme ve algılama sistemlerine ihtiyaç duyulmaktadır.

Bu çalışmada, dijital ortama aktarılmış Pap-smear örneklerindeki anamolileri belirleyebilen bir bilgisayar destekli analiz sisteminin geliştirilmesi amaçlanmıştır. Hücre çekirdeğinde büyüme ve sitoplazma alanlarında deformasyon gibi bozunmalar kanser varlığına ilişkin çok önemli ipuçları sağlayabilmektedir. Ancak Pap-smear örneklerinin bir çoğunda yaygın olarak görülen üst üste binmeler ve çeşitli gürültüler makinelerin doğru tanı koymalarını zorlaştırmaktadır. Bu nedenle geleneksel morfolojik görüntü işleme yöntemlerine ek olarak yeni bir segmentasyon yöntemi geliştirilmiş ve kullanılmıştır.

Önerilen segmentasyon yöntemi, akışkan dinamiğine dayanmaktadır. Hücrelerin yoğun biçimde üst üste bindiği durumlarda bile, modellenen akışkan, sitoplazma alanlarına nüfuz edebilmektedir. Bu sayede gerçekleştirilen bölütleme sonrası, hücre ve çekirdeklerden elde edilen sayısal özellikler sınıflandırılmıştır. Ayrıca çalışmamızda, geleneksel makine öğrenmesi yaklaşımlarından bazıları anormalite tespiti için denenmiş ve karşılaştırılmış, böylece pap-smear örneklerini tanılama becerisine sahip bir sistem ortaya konulmaya çalışılmıştır.

Anahtar Kelimeler: Pap-smear, *sınıflandırma*, Tanı, Akışkan dinamiği

ACKNOWLEDGEMENTS

I would like to express my gratitude to both of my supervisors, Assoc. Prof. Mutlu Avcı and Assoc. Prof. Mustafa Güven for their valuable suggestions and comments. I am highly thankful for their patience during the learning process of this master thesis.

I also wish to thank Instute of Natural and Applied Sciences and Scientific Research Projects Unit for all sources they provide during the preparation of this thesis.

Finally, I would like to thank my family, who have supported me throughout entire process, both by their love and encouragements.

CONTENTS **PAGE**

List Of Tables **Page**

LİST OF FİGURES **Page**

LIST OF ABBREVIATIONS

PAP **:** Papanikolaou

FCM **:** Fuzzy C-Means

RBF **:** Radial Basis Function

HDS **:** Herlev Data Set

1. INTRODUCTION

Pap Test, also known as Pap-smear (Papanikolaou) is a gynocological scanning test, depending on searching for changes in the cells on cervix (lower part of the uterus). Whole process roughly depends on the examination of the cells under the microscope, taken from the patient. Pap-test has a crucial role on early diagnosis of cervical cancer.

During 2003, about 12.200 women in the U.S. is diagnosed with cervical cancer, and 4,100 women died of the disease. (CDC, 2003) Furthermore Pap-Test is a very effective screening way for identifying cell abnormalities and infections therefore two times per year, regular Pap test is recommended for all women between the ages of 21 and 65 years old (CDC, 2013). Usually there are no symptoms caused by abnormal cells in the cervix especially at first. That's why pap smear is important for early diagnosis of cervial cancer. Cervical cancer is highly curable if it is found early.

Today, Pap test is usually completed by general practitioners, technicians and cytologs. Usually it is done in a office or clinic. Technician or nurse uses a plastic or metal speculum to widen the vagina during the test and few cells and mucus is collected from the cervix. These cells are sent to a laboratory on a slide for checking for abnormalites in cells (CDC, 2003). These abnormalities are detected via visual examination of cell samples due to their morphological and structural features. Identification of normal and abnormal cells during the test is achieved by these features such as relative size of cell nuclei, intensity and size of cytoplasm area and all unexpected abnormalities on color and morphology. All samples, seen under the microscope usually do not have the same characteristics. They slightly have general differences due to type of the pap test such as conventional pap or liquid based cytology. In spite of that differences; cells and abnormal cells, seen under the microscope have general common features and characteristics.

If we consider these common features and similarities which all samples have, we can conclude that pap test is extremely suitable for automation.

In most of the cases, substracting cell areas from background is an easy task for naked human eye because usually images of the samples have clear backgrounds and objects on that background have much more strong intensity. Also on the test images, other cell types such as epitel cells usually appear in characteristic colors and nuclei areas appear as darkest regions due to their density.

In a typical sample, several hundred thousand cells can be observed. Despite the mentioned clear and easy to segment appearences, most of the sample images have artifacts, overlapping cells, mucus, blood etc. These distortions on the samples increase the total time of entire screening. That is one of the reasons why desire to automate this screening has always been great (Dougherty, 2011).

Today, it is possible to achieve harder computational tasks in miliseconds with up to date hardwares. These new computational improvements give us ability to process digital image data faster and much more effective. Thats why more scientific researches have been actualised for automating pap-smear screening.

In this study, development of a computer aided Pap-smear diagnosis system is the primary aim. An algorithm for powering such a system should also be flexable and intelligent. During this study several different algorithmic approaches developed for overcoming potential difficulties.

The content of this thesis is arranged as given in the followings.

After introduction, in chapter 2, survey on automation of pap-smear screening literature is given. In the chapter 3, after a brief introduction to general structure of the developed algorithm, main processing steps and their results are presented. Several different test results of the developped system with different machine learning approaches disscussed in chapter 4. General performance of the system and possible improvements are disscussed in chapter 5.

2.LITERATURE SURVEY

2.1. Introduction

Since 1950, there have been numerous scientific studies for development of new methods to automate Pap-smear. One of the strongest motive that mostly lead scientist to study on that subject is crucial role of early diagnosis on curation of cervical cancer. When cervical cancer is found early and treated, it is highly curable (CDC, 2013). Cheap and fast methods to use Pap-smear screening may save many lives every year. In addition to that it is relatively easy to take samples from patients for Pap-smear test and it is possible to take many samples at once. All these reasons plus general characteristics of the test images cause incentive motivation for automation of pap smear screening.

In the 1950' s, first attempts for automation of Pap-smear test were generally based on size of the cancer cells. In the sample images cancer cells appears as larger and darker areas with a greater amount of stained material, than normal cells. With a basic tresholding algorithm these darker pixel groups were segmented in early studies. By the number of segmented pixels it is possible to score volume of cancer cells on sample images. One of the limitations of this method was the fixed thresholding; it was very sensitive to proper staining and proper system setup (Dougherty, 2011). Obviously these tresholding method fails when the cell overlap occurs and single treshold value causes the loss of flexibility of the system.

In the 1960' s Judith Prewitt studied for segmentation of red blood cells by analysing the histograms of the digitizied test images. Her method was also used with Pap-smear releated studies. According to her method, there are 3 peaks on the histogram of the sample images. These peaks indicate the tresholds of 3 different regions on the samples. Peak with the highest graylevel value indicates background, peak with the middle graylevel value indicates cell cytoplasm areas and the peak with the lowest graylevel value indicates cell nucleus. According to this presumption, darkest areas on the images are cell nucleus and brightest areas are the background. By that precalculation, treshold values change sample by sample and segmentation

task is completed with higher success. In spite of that much more adaptive approach, method was still very senstive to artefacts and distortion. In case of lack of homogenity on samples, overlapped cell groups and random dark areas easly manipulate histogram which causes failure with segmentation process.

In the late 1960s, a group at Toshiba, in Japan, started working on developing a Pap-smear screening machine they called CYBEST (Watanabe, 1974)

They used a differential histogram approach for the automated thresholding (Tanaka, Ikeda, Ueno, Watanabe, Imasato, 1977). This new method was basically supressing smooth areas while giving weight to edges on histogram. By that approach their method could segment sudden changes from smooth areas.

In this study gaussian gradient magnitude is used for detection of cell nuclei areas which is very similar to the method, used at CYBEST.

Same group is also used four features of the cells to achieve a correct classification of cancer cells. These features were: nuclear area, nuclear density, cytoplasmic area, and nuclear/cytoplasmic ratio. (Tanaka, Ikeda, Ueno, Watanabe, Imasato, 1977).

In addition to that, they have discovered some useful features such as chromatin pattern and the shape of the cell nuclei which may be very useful for classification but They couldn’ t reliably measure these features automatically because of technical disabilities on accurate focusing on nucleus. Nuclear shape was determined as the square of the boundary length divided by the surface area (Dougherty, 2011). Also in this study nuclear shape was one of the features, used for classification. In addition to that, correlation of nuclear boundary with an ideal circle is also used as one of the features.

During 1980’ s several groups have studied on automated Pap-smear screening systems similar to CYBEST. All was based on rare event model. In this model, analyses based on detection of few dark and large abnormal cells, among others. With this model, classification and segmentation should be fast and efficient for scoring correct cells. In many samples, overlapped areas and small clumps of various kinds makes this model impractical.

Because of that difficulties, several groups have developed a different approach to key features of cells. They' ve observed that, in the samples with cancer even normal cells have some changes at their features. These changes were not strong enough to classify a single cell but they were strong enough to detect abnormal specimens with statistical analysis. This phenomenon came to be known as MAC, malignancy associated changes (Burger, Jutting, Rodenacker, 1981)

The CYBEST group noticed that MAC phenomenon was very strong with cell nuclei chromatin and for deriving benefit from MAC it was necessary to extract nuclei chromatin but they have not been able to extract nuclei chromatin efficiently.

After conlusion of CYBEST group about chromatin pattern, A group at the British Colombia Cancer Research Centre developed a new and much more efficient algorithm for extracting cell nuclei chromatin pattern. Instead of using a single, global threshold to determine the nuclear boundary, group in British Colombia used the gradient magnitude to accurately place the object boundary (MacAulay, Palcic, 1990). They have used gradient magnitude to delineate the nuclei at first stage and then they have cleaned the artifacts which are touching the actual boundary of the cell nuclei. They have used grey level values of the pixels on the gradient magnitude image for a conditional erosion. According to that approach they' ve started to remove from low grey value pixels to higher ones. With that new method they were able to determine exact, clean boundaries of the cell nuclei for analysing chromatin pattern.

In the 1990' s, with the increasing commercial attempts, neural networks are used for the first time with studies on automation of Pap-smear. This new generation systems were ofrequently combination of morphological operations and artificial neural networks.

During this phase, cell nucleus was at the center of the studies. Segmented nuclei areas were analysed with artificial neural networks which are trained with many samples before. Classification of normal and abnormal cells is not a linear problem. Before neural networks it was very hard to cluster complex features with conventional linear algorithms. During these period, developed neural networks were working like a black box. Features which allow neural networks to determine

abnormal cells were unknown. These networks were capable to detect only few different cell appereances.

With the increasing process power of the computers black box model gradually lost popularity from the 2000' s. Various scientific and commercial groups have experimented with improved neural network approaches to perform better segmentation and classification. Thus, some of these groups have been able to get more successful results with different samples taken in different circumstance.

A group from Brisbane, Australia, has tried active contour model for the first time. According to that model, target areas are segmented roughly then these inital contours are moved theirselves to lower energy levels by each iteration. Energy level was calculated with 3 terms such as, pixel connectivity, curve of the pixels and and gradient magnitude of the pixel. Thus, contour was finding it' s way to accurate borders where the gradient magnitude value is higher, without loosing it' s closed form.

Different variations of this method is still used in segmentation of nucleus and cytoplasm altough method has shown low success with noisy areas or overlapping cells.

When we consider the work done from past to present about automation of Pap-smear test, we see two different groups of studies.

First group is focused on segmentation and feature extraction, second one is focused on classification of the extracted features.

Studies from these groups will briefly examined in following two sections

2.2. Studies Based on Feature Extraction and Segmentation

Many studies, focused on automation of Pap-smear screening, perform diagnostics cell by cell. This method of processing needs accurate segmentation of every single cell and nucleus for better classification. Thus, one of the common difficulties at these studies is detection of accurate boundaries before classification step. At the most of the studies, a number of features are extracted from previously

segmented cell boundaries for classifying abnormal and normal cells via different algorithmal approaches.

One of the expectations from the developed systems is ability of flexiblity during whole process. Which means, the percentage of success of the system should be conserved in all circumstances. Many studies have used artificial neural networks and machine learning algorithms for obtaining this flexiblity during classification step. It is obvious that, same machine learning approaches may be really helpful at segmentation stage.

In a study, published in 2003, a method based on enhancement of contrast is used for segmentation of pap smear cells. According to this approach, linear contrast enhancement and k-means clustering methods combined together for clustering pixels so that they were able segment cell and nucleus areas even with noisies and blur areas (Isa, Ashidi, Mashor, Mohd, Othman, 2003).

In another study, published in 2009, watershedding and active contour theorem is combined for accurate segmentation. Their algorithm was able to segment nuclei areas even with noisy areas and artifacts. (Kale, Aksoy, Önder, 2009)

As it seen from the previous studies focused on automation of Pap-smear, a combination of machine learning and morphological operations may be very useful for segmentation stage. Also another machine learning algorithms such as clustering methods instead of neural nets and cellular automata approaches may be efficient tools for accurate segmentation.

2.3. Studies Based on Classification and Diagnosis

As a conclusion form previous studies, it can be said that, classification of the segmented areas is the most crucial stage of developed systems. In most of the cases classification of abnormal cells is not a linear problem. Thats why conventional linear algorithms based on conditional branching are not enough for automated dignostics. Thus from past to present, machine learning algorithms find great place at every stage of studies related with automation of Pap-smear screening. By these machine algorithms, segmented pixel groups could be classified due to their key

features despite the all enviromental noisies and overlapped cell areas. These key features are mostly derived from global knowledge about cervical cells and common anomalies such as nucleus, cytoplasm area, nucleus perimeter, shape, intensity and size of nucleus.

In a study conducted at the University of North Carolina, 10 different features were used to be able to diagnose with neural networks and fuzzy logic based algorithm. The following features are used in the study: Percentage of cell coverage, cell coloration, distribution of cell size, average of cell size, existence of multinuclear cells, existence of halos in cells, distribution of nuclear size, average of nuclear size, nuclear size/cytoplasm size ratio, percentage of empty cells. These features were used for clustering samples into 2 classes (Li, Najarian, 2001).

In another study, conducted in 2006, 917 Pap-smear samples are examined in 7 clusters. 3 of them were belong to normal cells and 4 of them were belong to cells which have various anomalies. Following features were used for clustering samples in mentioned groups: Nucleus area, Cytoplasm area, N/C ratio, Nucleus brightness, Cytoplasm brightness, Nucleus shortest diameter, Nucleus longest diameter, Nucleus elongation, Nucleus roundness, Cytoplasm shortest diameter, Cytoplasm longest diameter, Cytoplasm elongation, Cytoplasm roundness, Nucleus perimeter, Cytoplasm perimeter, Nucleus position, Maxima in nucleus, Minima in nucleus, Maxima in cytoplasm, Minima in cytoplasm. These features taken place in classification of samples due to Mahalanobis distance norms. (Jantzen, Dounias, 2006).

3. MATERIAL AND METHOD

3. 1. Problem Statement

Pap-smear is one of the most common gynocological scanning tests and it is roughly based on searching abnormalities on the cervical cells under the microscope. A computer aided system for Pap-smear screening should anlayse the sample images in many stages and recognise abnormalities on cells. One of the most common abnormality that helps cytologs and automated systems to diagnose malignancy is abnormal growth of the nuclei. Also shape and density of the nuclei may shown as an example of some of the discrimination features. In this thesis, this kind of significant features were aimed to be extracted and used for classification.

There are three basic regions on the Pap-smear samples such as nucleus, cytoplasm and background areas which are shown in Figure 3.1.

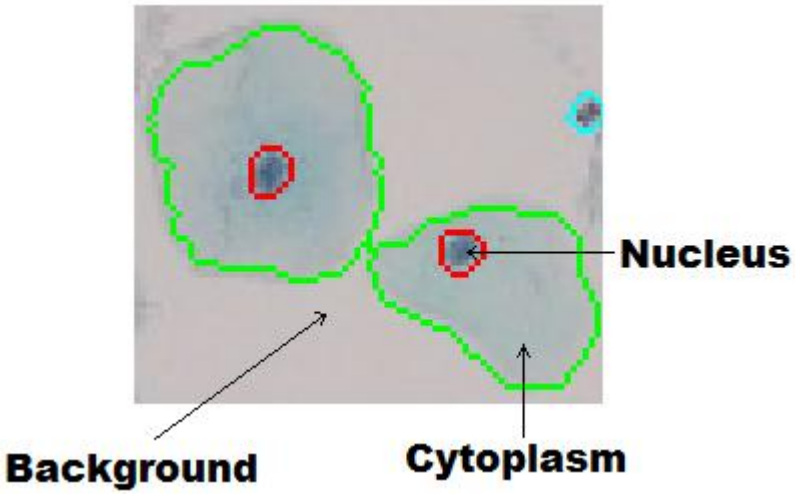


Figure 3.1. Three regions of a segmented cervical cell on a Pap-smear sample.

A computer based automated diagnosis system should be capable of segment both nucleus and the cytoplasm regions of the cells from digital images to use mentioned features for ranking.

Simple tasks for human eye such as distinguishing cells from background may be extreme hard for computer vision systems. Unlike human eye, digital screening systems easly mistaken by small differences or distortions on the samples.

Methods, proposed in this thesis were optimised for maximum success rate with multiple samples.

On most of the Pap-smear test images, there are nested regions where cell overlapping occurs. An example to nested areas from one of the samples is shown in Figure 3.2.

Expectation from a succesful screening system for malignancy is capability of recognising adjacent areas. Such a system should be able to determine locations of both nuclei and cytoplasm regions even when images with high degree of overlapping are used.

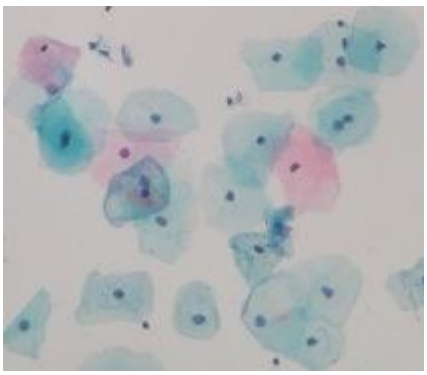


Figure 3.2. An example of nested regions from one of the samples

Obviously classification of the segmented cervical cells is not a linear problem. Which means it is not possible to rank abnormality of the segmented cells with conventional linear algorithms. This kind of non-linear problem may be solved by several types of machine learning algorithms. Some of them requires pre-classified training sample sets. They basically learn the features of the different classes then try to decide. Also some of them doesn’ t need a trainer for classification. Trained or not, developed system for cervical cell diagnosis should have a flexable decision mechanism which is capable of recognise classes even with missing or varying features.

3. 2. Pap-smear Sample Images and Attributes

Two groups of samples were examined in this study. First group was consist of images from Herlev Data Set and used for training artificial neural network. HDS is formed by Department of Pathology at Herlev University Hospital and the Department of Automation at Technical University of Denmark and it is collected for providing benchmark data for classification experiments.

All images in the set are segmented single cells which are collected by cyto-technicians. Each of the images are taken by a digital camera connected to microscope with a magnification of 0.201 *µm/pixel*. Segmentation of cells are completed by a commercial analyses system named in CHAMP (byriel, 1999). A segmented cell by CHAMP and it' s different regions are shown in Figure 3.3.

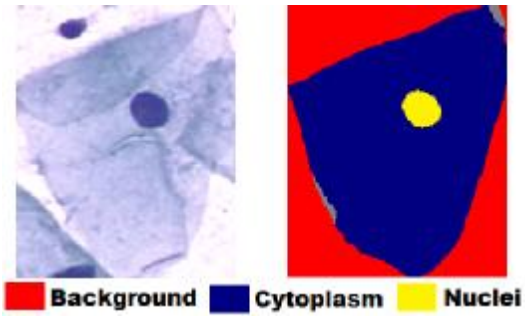


Figure 3.3. One of the segmented cell samples from Herlev Data Set.

There are seven groups of images in the HDS. Each of these groups are representing different malignancy classes. These classes were created by cyto-technicians manually by examining the segmented cells one by one. Previously classified cell images are used for training machine learning algorithms in this study. Different classes of HDS are shown in Table 3.1. and Table 3.2.

Table 3.1. Three normal classes of Herlev Data Set and some of their features.

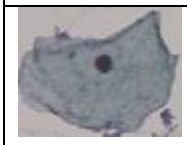
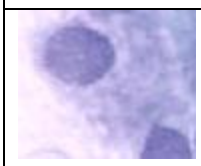
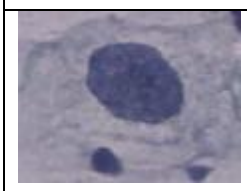
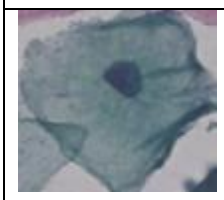
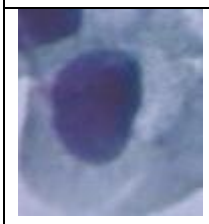
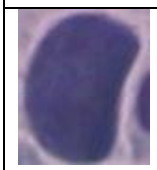
Classes Of Normal Cells		
	Name of the Class:	**Superficial Squamous**
	Nucleus Cytoplasm Ratio:	Very Small
	Nucleus Radius:	Small
	Name of the Class:	**İntermediate Squamous**
	Nucleus Cytoplasm Ratio:	Small
	Nucleus Radius:	Grown
	Name of the Class:	**Columnar**
	Nucleus Cytoplasm Ratio:	Medium
	Nucleus Radius:	Grown

Table 3.2. Four abnormal classes of Herlev Data Set and some of their features.

Classes Of Abnormal Cells		
	Name of the Class:	**Light Dysplasia**
	Nucleus Cytoplasm Ratio:	Medium
	Nucleus Radius:	Large
	Nucleus Density:	Low
	Name of the Class:	**Moderate dysplasia**
	Nucleus Cytoplasm Ratio:	Large
	Nucleus Radius:	Large
	Nucleus Density:	High
	Name of the Class:	**Severe dysplasia**
	Nucleus Shape:	Deformed
	Nucleus Radius:	Large
	Nucleus Cytoplasm Ratio:	Very large
	Name of the Class	**Carcinoma in situ**
	Nucleus Shape:	Deformed
	Nucleus Radius:	Very Large

Other set of samples used in this study were consist of random test images obtained from several experimental sample sets which are compiled for scientific purposes. Observation frequency of artifacts and distortions on choosen samples were similiar to regular Pap-smear samples. Testing developed algorithm with regular samples is very crucial for the objectivity of the developed methods. There are several extreme examples in the sample set for observing the behaviour of the code with different circumstances. Every stage of the developed code was tested with different samples even with extreme ones for observing how system reacts stage by stage.

Pap-smear sample images in this set were stored in JPEG format with various sizes. Two examples from set are shown in Figure 3.4.

Samples, bigger then 640x480 were reduced for speeding the process up. Also, focal lenght and zoom ratio is fixed with all samples. Thus scale of the cells was same at the all samples.

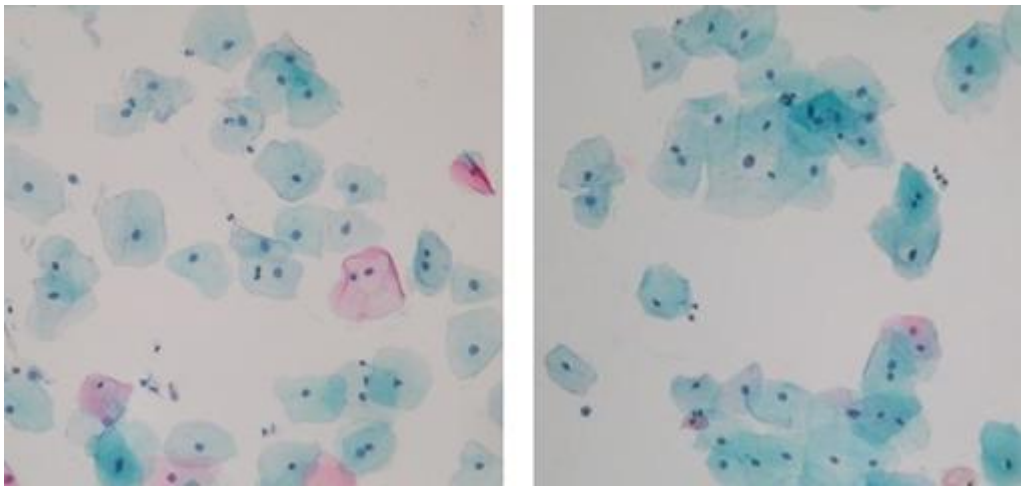


Figure 3.4. Two regions from some of the samples, used in study.

In most of the samples cell and nucleus boundaries were distinct and clear enough for diagnose via naked human eye but also, there were overlapped cells and artifacts in almost all of the sample images.

Generally two types of cells appear on the samples. Cells, seem in shades of red are blood cells and seem in shades of blue is cervical ones which are desired to be examined.

Homogeneous distribution of background pixels makes the background segmentation easier in most of the cases. However, various artifacts and irregularities which are unnoticeable for naked eye, also complicate the segmentation with linear decision algorithms and morphological opearations. Some of these factors are shown in Figure 3.5.

It is obvious that an adaptive, intelligent decision mechanism is a necessity for a practical diagnosis system.

(a) (b) (c)

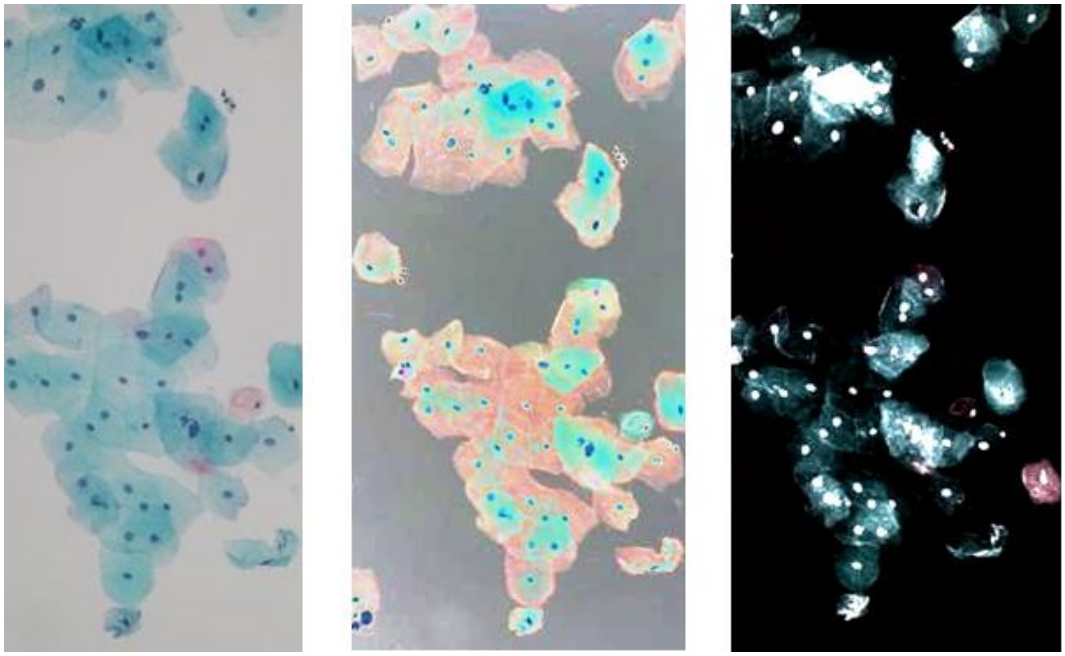


Figure 3.5. (a) A region from one of the samples with overlapped cells and various artifacts. (b) Same region with manipulated luminosity. (c) Same region with manipulated RGB values

As it' s seen on the Figure 3.5, it was not possible for human eye to notice every artifact and noisy on samples before pixel value manipulations. These unnoticable distortions cause failure of computers (also numerical computations) on classification and segmentation stages.

3.3. Structure Of The Algorithm

MATLAB is prefered as development enviroment in this study. However, at the several stages of the code, algorithm is tested with c++ under both linux and windows as a practical application for effective usage of system resources.

Also, OpenCV library which is developed by intel took place for basic operations in study for saving time. It is one of the common libraries frequently used in professional image processing applications.

Developed algorithm during the study had 6 basic stages. These are, preprocessing, segmentation of cells from background, determination of the locations of nuclei in the image, determination of cell boundaries by fluid segmentation engine, feature extraction and classification. Block diagram of the proposed system is shown in Figure 3.6.

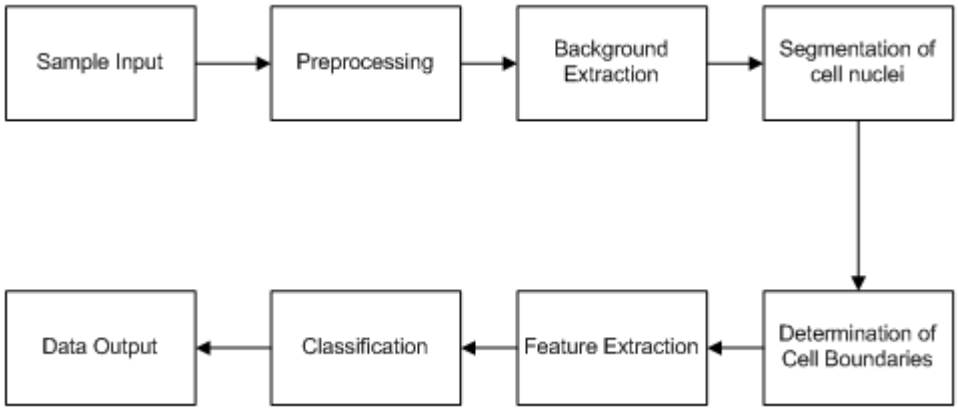


Figure 3.6. General structure of the algorithm

3.4. Preprocessing

Preprocessing step aims to improve the general characteristics of samples for increasing the success before any analysis. Also raw samples reduced to size of 640x480 if they were not sized appropriately and some basic filters were applied for sharpening the edges.

3.5. Background Extraction

Before any diagnosis on cells, all the objects on background should be segmented from image. Accordingly, in this step all pixels were regrouped in two general clusters as background pixels and object pixels. Object pixels include all cell cytoplasm areas, artifacts and other detectable information in it.

After the elimination of background pixels, remaining pixel group will cluster again for detection of cell and nuclei boundaires for classification at the next stages.

As it' s seen on the sample images, cell and nucleus boundaries usually appear sharp and clear and there were apparent differences between background and other pixels. Based on that facts, it is envisaged in this stage of the study, segmentation of background pixels could be handled as a linear problem with various tresholding methods. It is presumed that, applied tresholding will work with similiar success at the most of the samples. Obviously conservation of success depends on similiarity between the photographic standards of the samples. In most of the samples, there are plenty of distinctive features which may allow us segment background pixels from others. Hue value is used as a classification feature in this step of the algorithm. According to that, samples are converted to HSV from RGB format then background pixels are segmented due to their hue value.

Two treshold values were determined for this process, one for upper limit and one for lower limit. Same treshold values were used for all samples during the study. All pixels in this band were presumed cell area pixels. In most of the hue histograms of the samples, there are three peaks. Middle peak carries the pixel values of the cytoplasm areas. According to that information treshold values were choosen carefully for maximum performance with all sample images. Hue histogram of one of the samples with tresholds on it is shown in Figure 3.7.

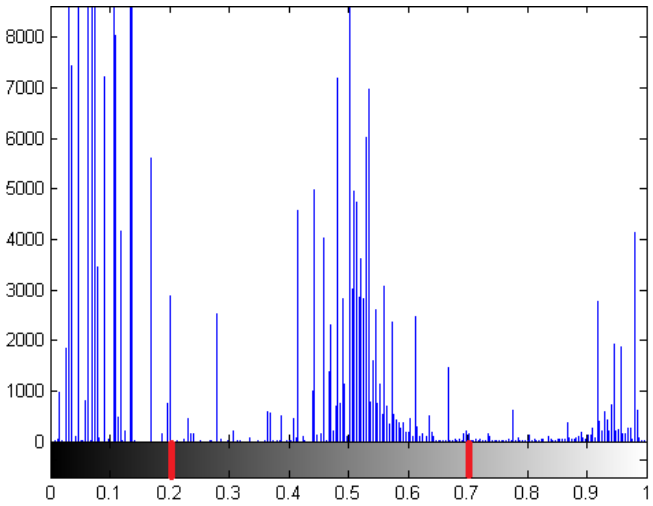


Figure 3.7. Hue histogram of one of the samples with determined upper and lower tresholds.

After tresholding, all connected pixel groups were labelled with morphological operations and some small objects were removed. Finally perimeter of the remaining connected area was labelled as total cell area. Outputs of process steps for two sample regions are shown in Figure 3.8.

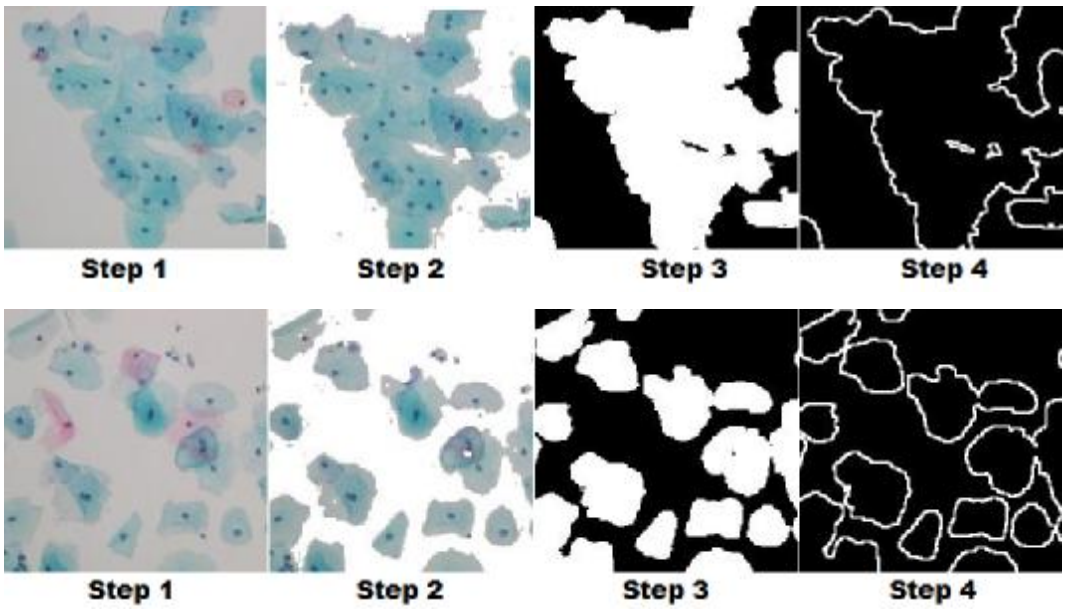


Figure 3.8. Outputs of segmentation process steps for two sample regions.

3.6. Segmentation Of Cell Nucleus

Cell nuclei appears as one of the darkest regions on most of the samples. There are also other darker regions such as areas of artifacts, groups of cell nucleus and several types of noisies. Some of the previous studies were constructed on that global data which allows us to assume place of cell nuclei is in the lower grey level areas.

A bunch of previous studies have accomplished nuclei segmentation with single or multiple stage tresholding. Tresholding is a linear decision mechanism which may cause data loss on segmented regions. It is crucial to determine accurate nuclei boundaries for future extraction. Even with adaptive methods, tresholding was insufficient for practical applications when used alone. Some of the sample images before and after tresholding are shown in Figure 3.9.

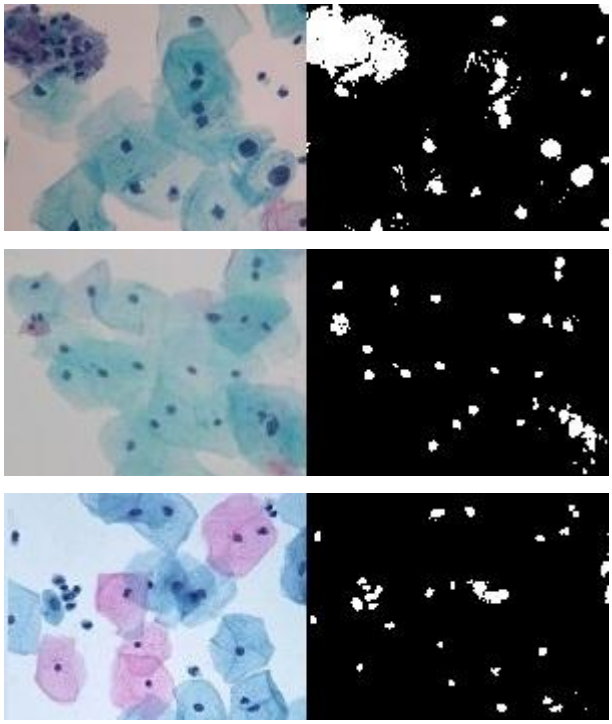


Figure 3.9. Some thresholded sections from several samples

As seen on the figures, a linear decision system does not conserve it' s efficiency at different samples or regions. Tresholding may be a solution for

background segmentation but it is clearly insufficiant for detecting nucleus regions precisely. Also tresholding methods mostly cause data loss or unwanted pixels at the segmented regions. To overcome these disadvantages, gradient magnitude of the samples were used for nucleus segmentation during the study Segmentation process roughly have two main stages. At the first stage gradient magnitude of the image was extracted then it was filtered with edge detection filter. Finally, filtered black and white image was cleared and locations of the cell nuclei were much more accurately determined.

3.6.1. Gradient Magnitude

Gradient magnitude of a digital image is basically accentuation of regions where sudden changes are observed on the grey level of the image. According to that, sudden change regions, mostly object boundaries, will be more apparent after the gradient magnitude operator applied (Jacobs, 2005).

3.6.1.1. Calculation of Gradient Magnitude

It is possible to form gradient vector as combination of the partial derivative of the image in the x and y direction. Which we can express gradient vector in the following form,

$$\nabla I = \left(\frac{\partial \mathrm{I}}{\partial \mathrm{x}}, \frac{\partial \mathrm{I}}{\partial \mathrm{y}} \right) \qquad (3.1.)$$

In this equation partial derivative indicates how rapidly image intensity changes as x and y changes.

For calculating partial derivative of the image matrix first we should consider derivative of a continious function I(x,y), which we can express as in (3.2.)

$$\frac{\partial I(x,y)}{\partial x} = \lim_{\Delta x \to 0} \frac{I(x+\Delta x, y) - I(x,y)}{\Delta x} \tag{3.2.}$$

In this case, derivative is taken at only one pixel intervals. So that we can express the same equation for both x and y as following:

$$\frac{\partial I(x,y)}{\partial x} \approx \frac{I(x+1,y) - I(x-1,y)}{2} \tag{3.3.}$$

$$\frac{\partial I(x,y)}{\partial y} \approx \frac{I(x,y+1) - I(x,y+1)}{2} \tag{3.4.}$$

Sobel filter is used as discrete differentiation operator for applying this operation to the whole image during the process. Sobel filter is based on vertical and horizontal convolution of image matrix with two 3x3 kernels. If we consifer A as the source image,

$$\mathbf{G}_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * \mathbf{A} \tag{3.5.}$$

$$\mathbf{G}_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix} * \mathbf{A} \tag{3.6.}$$

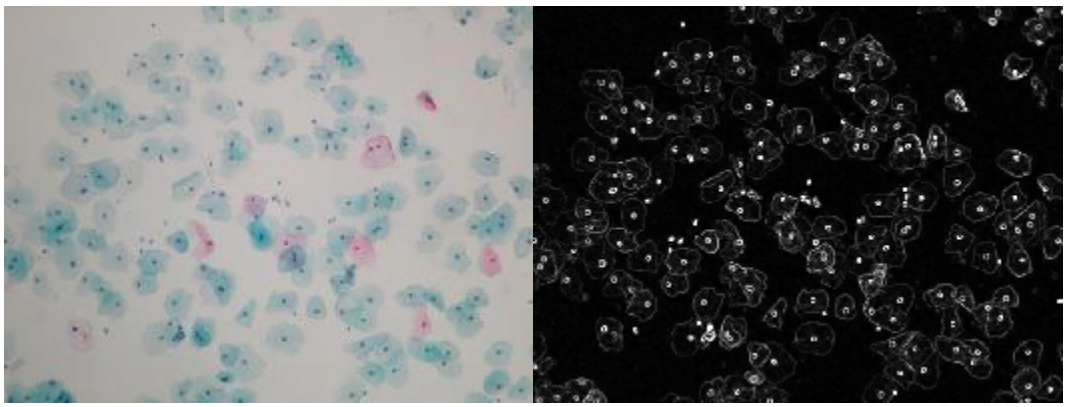
This equations generates two different images as Gx and Gy and " *" stands for two dimensional convolution. These images contain the vertical and horizontal derivative information.

Finally, gradient magnitude image is generated by following equation:

$$G = \sqrt{(G_x{}^2 + G_y{}^2)} \qquad (3.7.)$$

Gradient image of one of the samples is shown in Figure 3.10.

(a) (b)



(c)

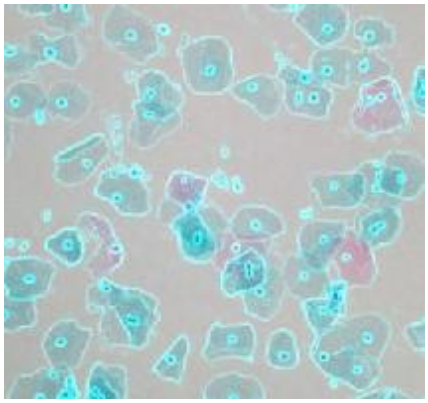


Figure 3.10. (a) Original image (b) Gradient magnitude image (c) Output image with accentuated edges by adding gradient magnitude data to the actual image

As it' s seen on the sample outputs, cell nuclei and cytoplasm area contours were accentuated and some of the artifact areas were suppressed at gradient magnitude images.

3.6.2. Filtering

After the extraction of gradient magnitude image, cytoplasm and nuclei boundaries are much more appearent as edges as a result of density differences between these areas.

However even with gradient magnitude data, it was still not possible to segment nuclei and cytoplasm regions with a linear decision mechanism and there were still artifacts and distortions on the final image. These unwanted objects mostly appear as edges and small pixel groups with higher gray levels. An example of a distorted area from a sample is shown in Figure 3.11.

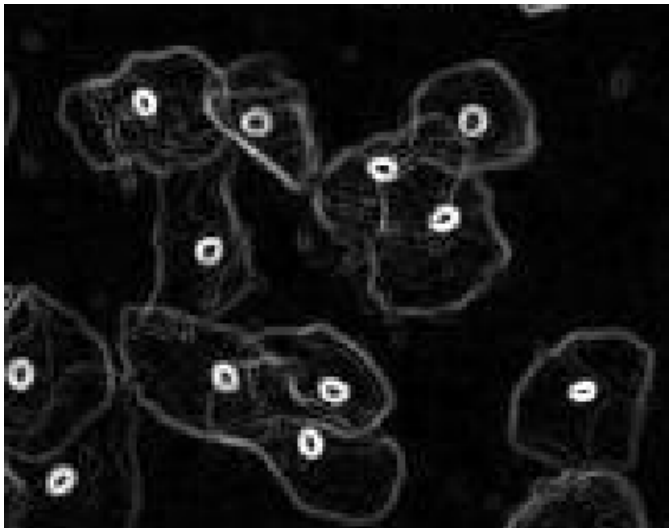


Figure 3.11. A particular area with artifacts and unwanted data from gradient magnitude image of one of the samples.

For eliminating these remaining pixel groups and obtaining cell and nuclei contour data, images were filtered with canny edge detection filter in the first step.

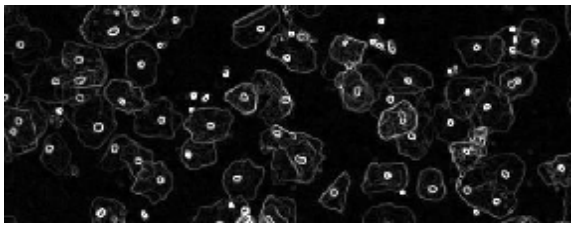
Canny filter simply completes these series of steps for edge detection:
Image is blurred for noise reduction by convolving the image with gaussian filter with given sigma value (Kalra, 2009). In (3.8.) an example of a 5x5 gaussian filter with sigma 1.4 is shown.

$$B = \frac{1}{159} \cdot \begin{bmatrix} 2 & 4 & 5 & 4 & 2 \\ 4 & 9 & 12 & 9 & 4 \\ 5 & 12 & 15 & 12 & 5 \\ 4 & 9 & 12 & 9 & 4 \\ 2 & 4 & 5 & 4 & 2 \end{bmatrix} \quad (3.8.)$$

Then gradient magnitude image is extracted and non maximum supression is applied to the final gradient magnitude image. Aim of this last step is supressing the non local maximum pixel which means, preserving only local maxima pixels and removing all of other pixels. And finally hysteresis operator is applied. According to this last step, every pixel is assigned as one of three labels by a treshold value. Pixel is a strong edge pixel, pixel is a weak edge pixel or pixel is not an edge pixel. Weak edge pixels which are touching to the strong edge ones are labelled as strong edge pixels. Finally, only strong edge pixels are valued as 1 and others valued as 0 so final image is black and white and only strongest edges are still on the image.

At that stage of this study, recent image was filtered with two different max hysteresis treshold value. First one is for extracting nucleus areas other one is for cytoplasm areas. Necessity for two different treshold values, comes from the intensity differences of nucleus and cytoplasm areas. Treshold values are determined empirically. Trashold is determined as 0.4 for nuclues and 0.04 for cytoplasm areas. A particular area after filtered with different tresholds is shown in Figure 3.12.

(a)



(b) (c)

Figure 3.12. (a) Gradient magnitude image (b) Filtered gradient magnitude image with 0.4 max. Hysteresis treshold (c) Filtered gradient magnitude image with 0.04 max. Hysteresis treshold

After canny edge detection filter two types of output data retrieved. Data that includes cytoplasm contours will be used at the next stages of the algorithm. A region with both extracted cytoplasm and nucleus boundary data is shown in Figure 3.13.

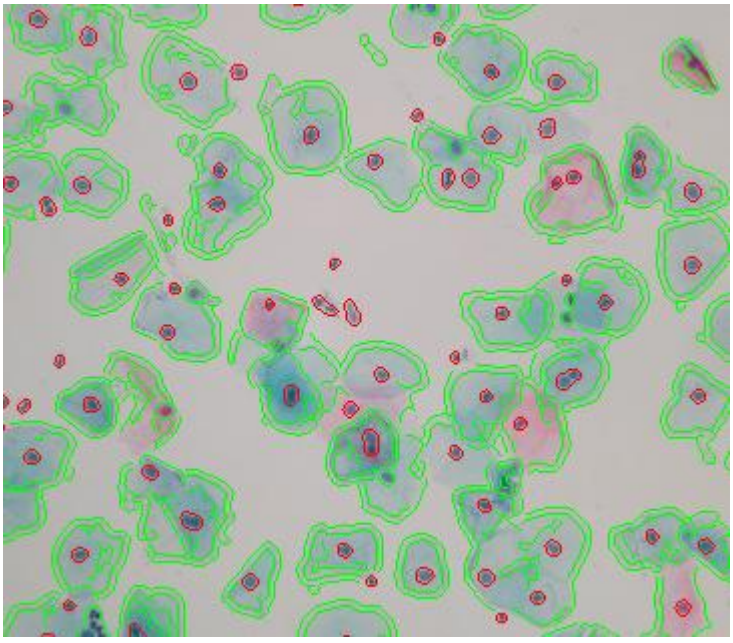


Figure 3.13. A part of an sample image marked with edges retrieved via two different hysteresis treshold.

As it' s seen on the Figure 3.13, output image, which is extracted with 0.4 max. Hysteresis treshold from gradient magnitude of the actual image has almost all of the nuclei boundaries in it, but it also still contains noisies and unwanted pixel groups. This final image has several types of connected objects: Nuclei boundaries (mostly appear as circular connected pixel groups), remaining strong edge pixels which belong to cell walls and artefacts. For filtering nuclei boundaries from this various data collection several types of morphological operations combined together.

At the first step of the process, all connected pixel groups are labelled. Then small pixel groups which can not be nuclei boundary are eliminated due to their size.

After small object removal, all connected objects were 2d correlated with an ideal circle which has same circumference with object. Low correalation value means object is not a circular group of pixels. So, it is possible to detect connected closed contour objects. A sample area after cleaning process is shown in Figure 3.14.

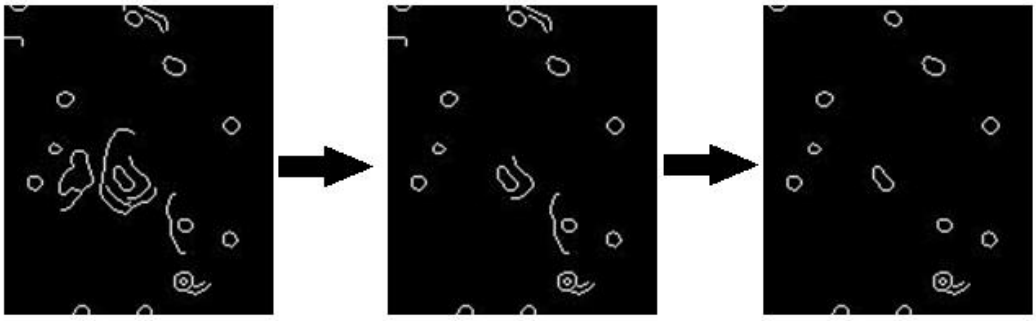


Figure 3.14. (a) Filtered raw image (b) Some of the objects are removed from image due to their sizes (c) Low correlation valued objects are removed.

Some of the remaining connected objects have small pixel groups inside them. It is possible to detect and remove these objects by filling insides of the all connected objects and substracting actual image from filled one. Result will be insider pixels of the all connected objects. Substraction of these pixels from actual image again will remove the all insider pixel groups.

Recently, nucleus walls were detected as closed contour connected pixel groups but there were still unwanted extra pixels around the walls which were

touching to wall pixels. These bifurcations should be removed before feature extraction. To achieve this final elimination all neighborhoods of each pixels of the object was checked and if a pixel doesn' t touch to insider pixels of the object then it is removed. Process is visualised in Figure 3.15.

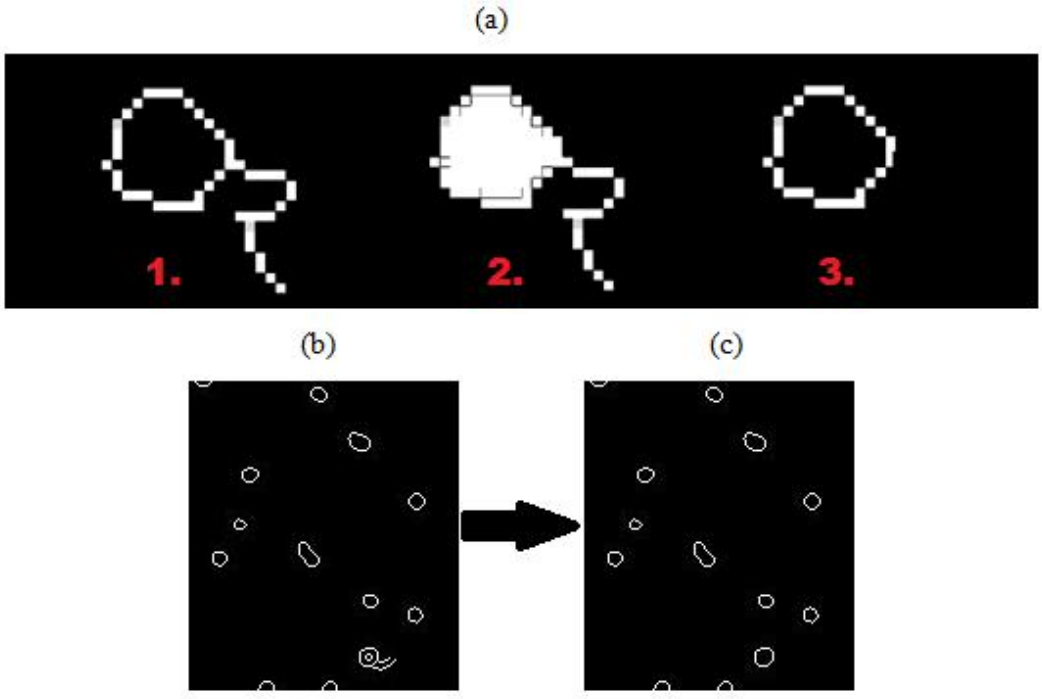


Figure 3.15. (a)An example of bifurcation elimination process (b)A region before final elimination (c) Same region after final elimination

At that stage of the study, only combination of morphological operations were used for nuclei boundary extraction.

It was mostly possible to retrieve this data without non-linear decision algorithms because of general characteristics of cell nuclei appearence on Pap-smear sample images. Two sample regions with detected nuclei walls are shown in Figure 3.16.

However, it is obivous that, morphological operations alone are not enough for complete segmentation of cells at practical applications. At the next stages of this study, some nature inspired algorithms will be combined with pixel based operations for cytoplasm segmentation.

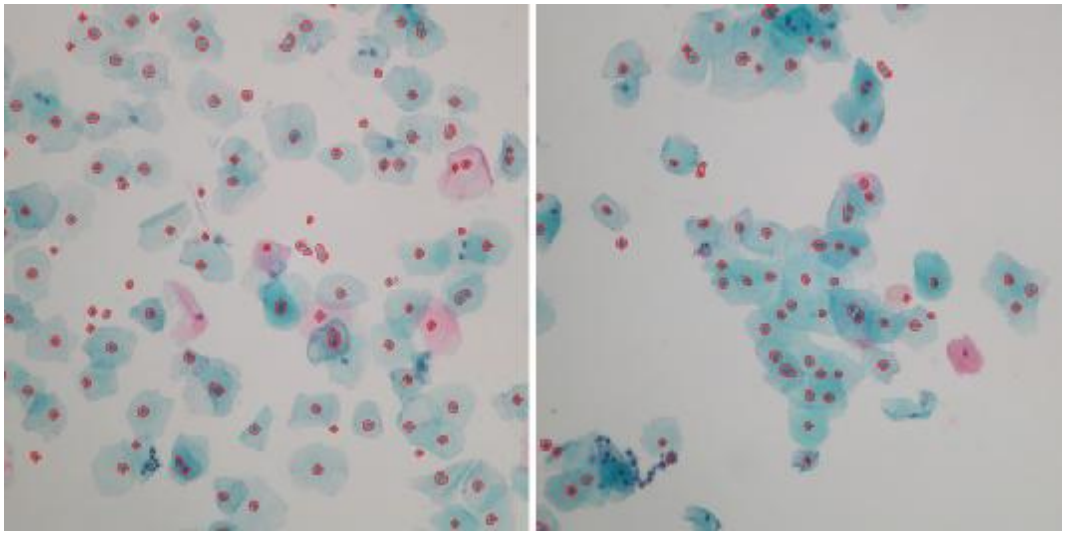


Figure 3.16. Dedected nucleus walls at two sample regions

As it' s seen on the Figure 3.16, efficiency of the developed systematic for nuclei detection depends on quality of the image and also locations of the cell nuclei. spatially closer cell nuclei may cause misclassification of the detected objects.

Nuclei closer to each other which are perceived as a single object is a critical problem. Even with clean and clear samples, algorithm detects some of the adjacent cells as a single object.

One of the most significant feature that indicates malignancy on the cells is size of the cell nuclei. If two or more adjacent nuclei labelled as one big object, this may cause misclassifications of the detected objects so that, additional analyses were applied to samples for detecting adjacent nucleus. Several types of features were used to analyse and detect adjacent nuclei with clustering methods.

.

3.6.3. Feature Extraction From Detected Cell Nuclei

Feature extraction from cell nuclei is a crucial step for the sake of completing all process with high success rate. In this study, feature extraction from cell was completed in two steps, at first, features of the nuclei then features of the cytoplasm were extracted and cells on the samples were classified by these features via machine learning algorithms. Also adjacent nucleus are detected and labelled by these features. All of the extracted features are shown in Table 3.3.

Table 3.3. Features, extracted from detected cell nuclei regions.

1	**Mean Intensity:**	Mean of the the grey level values of the all pixels in the region.
2	**Maximum Intensity:**	Maximum intensity value in the region.
3	**Minimum Intensity:**	Minimum intensity value in the region.
4	**Area:**	Area of the region as number of pixels.
5	**Convex Area:**	Area of the convex region as number of pixels.
6	**Equivalent Diameter:**	Diameter of a circle with same area as region as number of pixels.
7	**Extent:**	Ratio of total bounding box and area covered by region in the total bounding box.
8	**Filled Area:**	Area of the region as number of pixels after all holes are filled.
9	**Eccentricity:**	Scalar value that indicates how much is the region ellipsoid.
10	**LocalMin Number:**	Number of the locally minimum pixels.
11	**LocalMin Distance:**	Total distance between locally minimum pixels.

3.6.4. Detection of Adjacent Cell Nuclues

After the morphological segmentation of cell nuclei process, there were still unlaballed nuclei-like objects on the sample images. Some of these objects were various types of distortion and at the same time, there were still actual nuclei region data in these unlabelled information. In this study that remaining data was labelled as unidentifed objects for future operator review.

Without these unidentified objects there were still some misdetected regions, labelled as single cell nuclei which were closer to each other and perceived as a single object. Detection of these adjacent cell nuclei is necessary for the sake of the success of whole classification system. An example area with adjacent nucleus on it is shown in Figure 3.17.

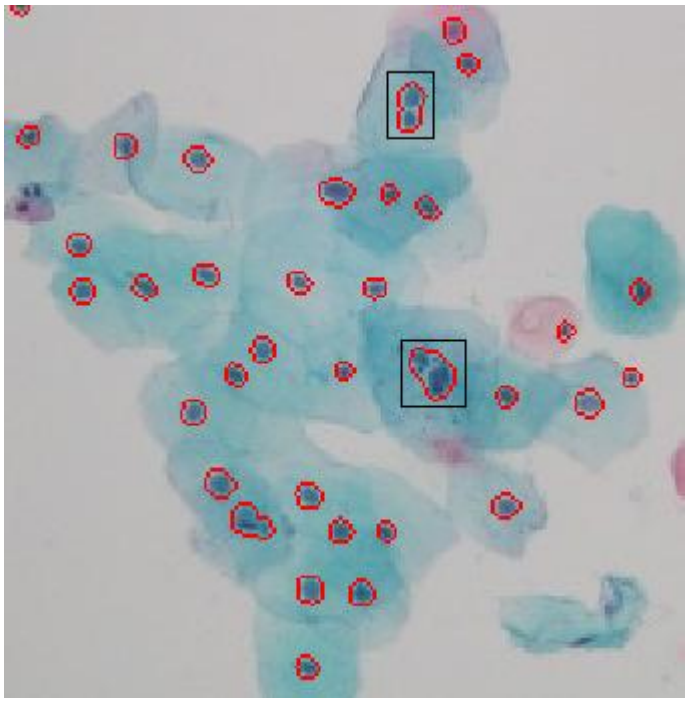


Figure 3.17. Some of the segmented adjacent regions are black boxed on a sample image

For detecting these adjacent regions, some of the previously extracted features were used. Picked up features were used for clustering all detected objects in two main classes as normal and adjacent regions. Following are the choosen features

for that clustering: Area, EquivDiameter, Eccentricity, LocalMin Number, LocalMin Distance.

These features are choosen carefully for seperating adjacent nuclei from others. There might be also possible abnormal nuclei inside the detected obejcts, these regions may be larger and misformed. These abnormal areas will be analysed at the next stages. So that, Optimum features were required for distinguishing abnormal and adjacent nuclei regions. Number of the Local minimum pixels and distance between these pixels are two of the most discrete features of the adjacent regions from others.

Locally minimum pixels were the lowest grayscale value pixels in the region and they mostly indicate the highest density on the region. In most of the adjacent nuclei cases, there were more than one local minimum point and euclidean distance between these points were bigger than one pixel. Local minimum differences between normal and adjacent nucleus areas are shown in Figure 3.18.

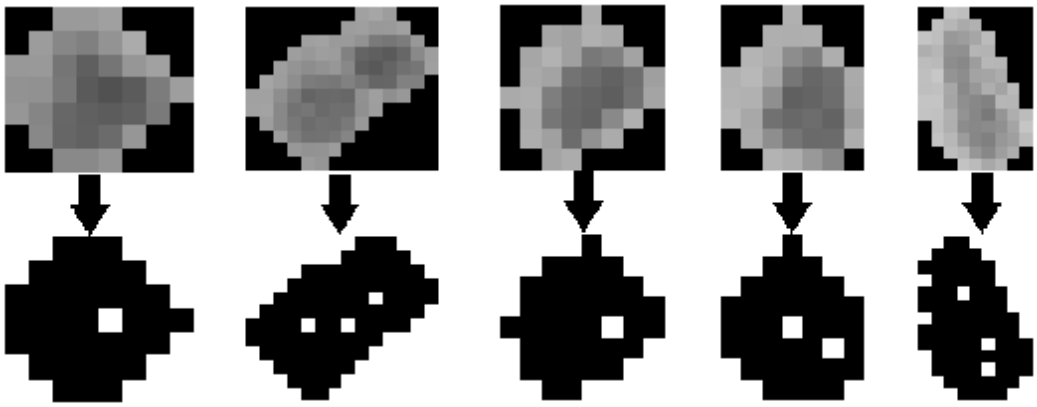


Figure 3.18. Local minimum points of some of the segmented regions.

Obivously, choosen features were highly discrete but even with suitable features, a smart and fast non-linear decision system is still necessary for succesfully classifying adjacent regions. In this thesis study two different methods were examined and compared for better results. These methods are mentioned below.

3.6.4.1. K-Means Clustering Algorithm

K-Means clustering method is based on relocation of " n" observations in a sample space. Samples supposed to be closer to the nearest centered group after every iteration. Clusters are formed by observations which have similiar features. Center of a group is mean of the all observations in the group (Kardi, 2007). Distance of an observation to centroids is recalculated with every iteration so that it is possible to cluster all similar observation into " n" clusters.

Computational process is carried out by a method named in Lloyd' s algorithm (also known as Voronoi iteration). According to that method, first input points are divided into K number of initial sets. Then method calculates the centroids and assign each observation to the closest centroid (3.9)

$$S_i^{(t)} - \{x_p : \|x_p - m_i^{(t)}\| \leq \|x_p - m_j^{(t)}\| \ \forall \ 1 \leq j \leq k\}, \qquad (3.9)$$

Usually " Euclidean Distance" is used for calculation of distance between points and centroids. After the association of all points to the closest mean, centroids are relocated (3.10).

$$\mathbf{m}_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\mathbf{x}_j \in S_i^{(t)}} \mathbf{x}_j \qquad (3.10)$$

These two steps are repeated by algorithm until centroids are fixed. (Tan, Steinbach, Kumar, Ghosh, 2006) At the end of the iterations all observations are divided into " k" number of clusters. Steps of k-means clustering is shown in Figure 3.19. Centroids of the clusters after process is shown in Figure 3.20.

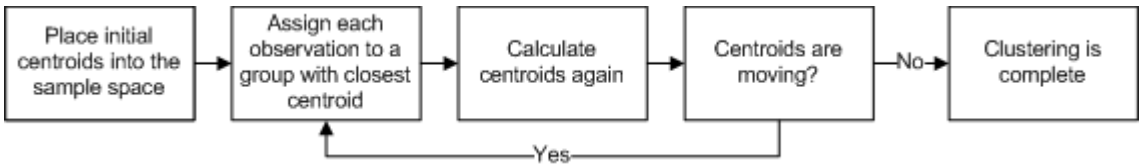


Figure 3.19. Block diagram of k-means clustering method.

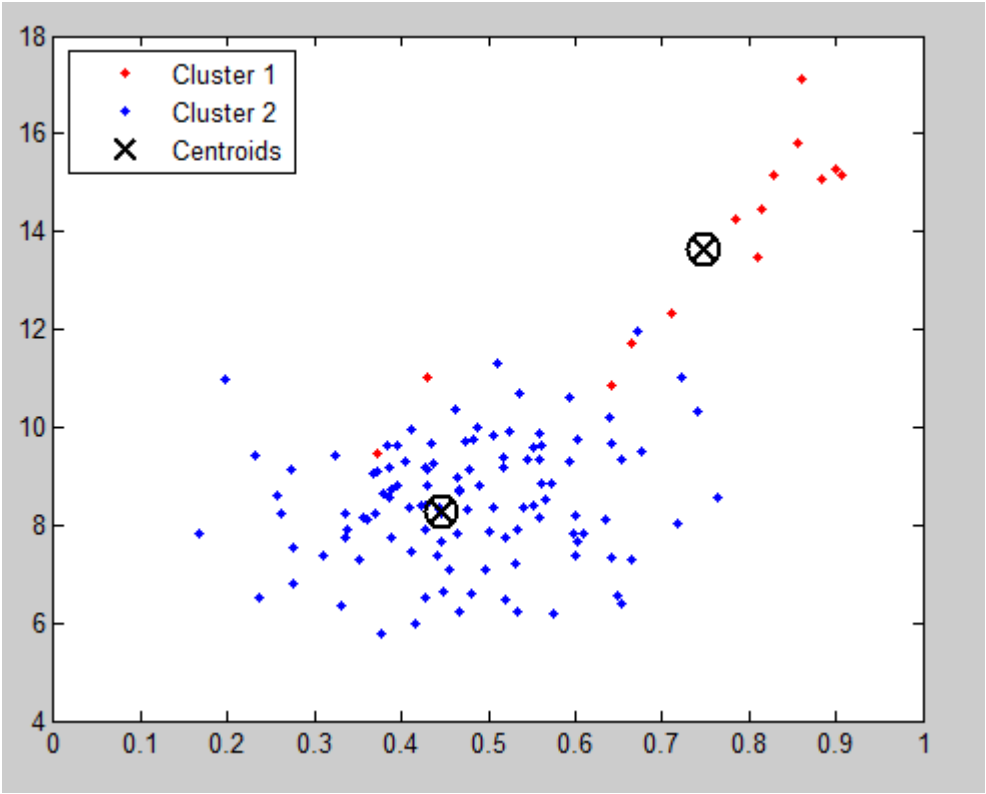


Figure 3.20. Clusters and cluster centers of a set of cells from one of the sample images

3.6.4.2. Fuzzy C-Means Clustering Algorithm

Fuzzy c-means clustering has similiar functionality with k-means but in fuzzy clustering every observation has a association degree to all sets according to their distances and observations does not belong to just one cluster. All points have a set of ceofficients (Hathaway, 2001). Each of these coefficients are degree of association with one of the clusters and Centroids of clusters are the weighted means of the sets. Which is shown in (3.11).

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{3.11}$$

Degree of belonging of a point to a cluster is inverse of distance between centroid of the cluster and the point so bigger distance means, lower degree of belonging to a set which is shown in (3.12.)

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{3.12.}$$

Centers of the clusters and coefficients of the points should be recalculated with each iteration until the change of coefficients are less than the given treshold. Block diagram of the process is shown in Figure 3.21 and clusters from a sample after process is shown in Figure 3.22.

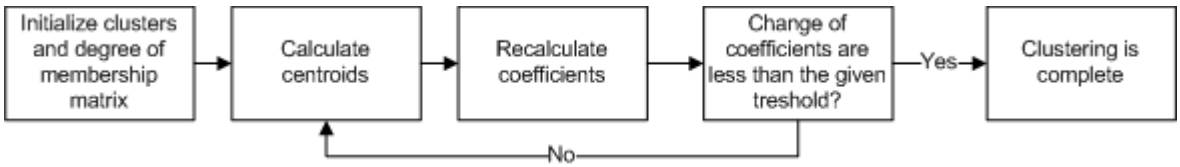


Figure 3.21. Block diagram of c-means clustering method

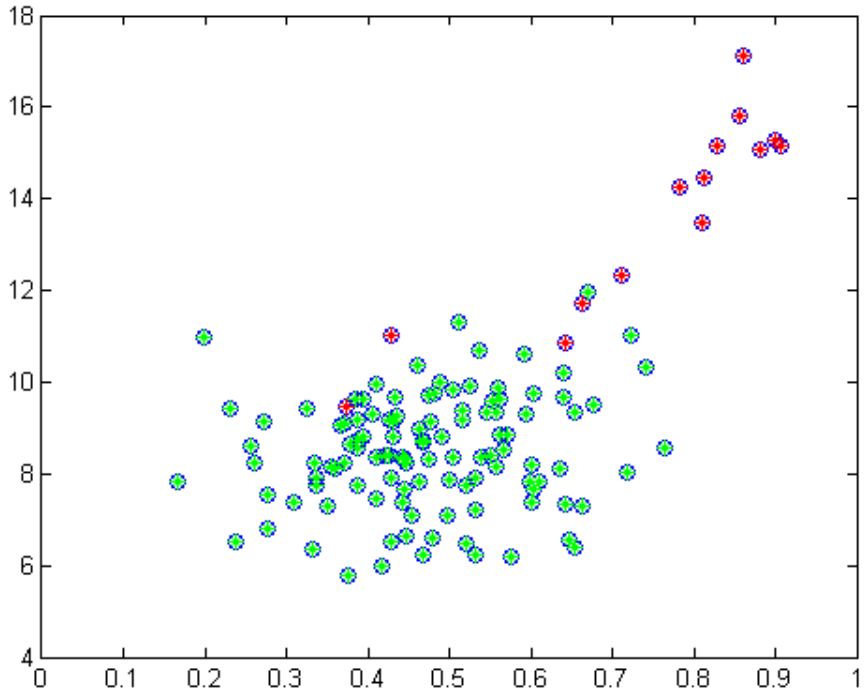


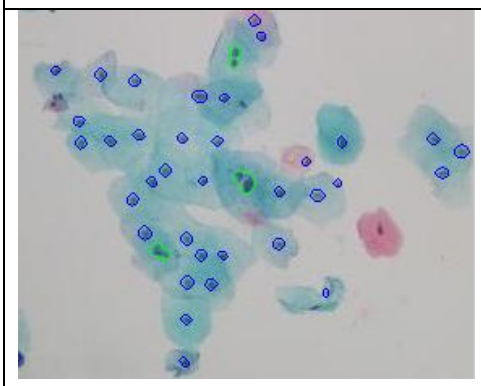
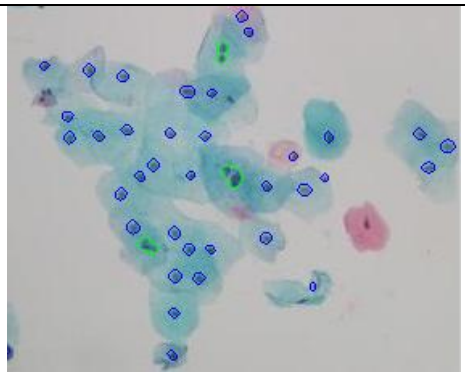
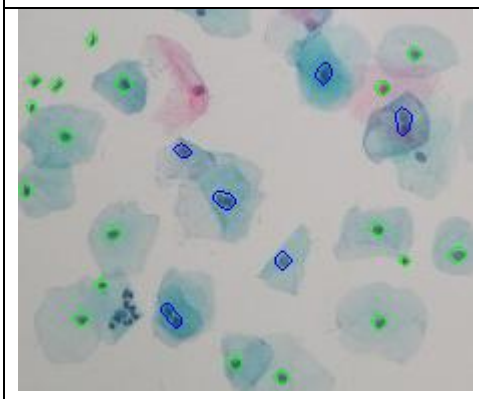
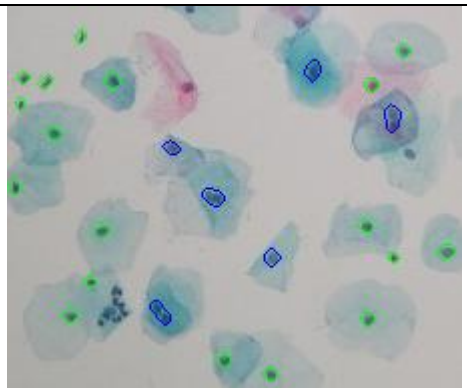
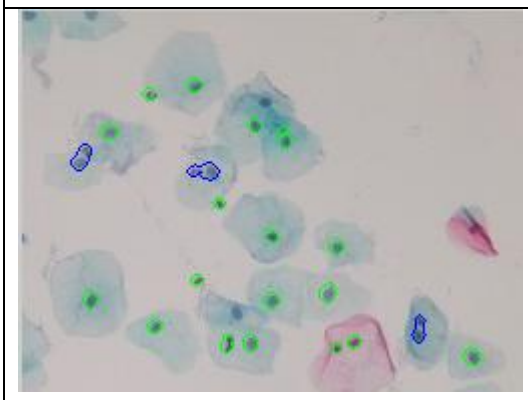
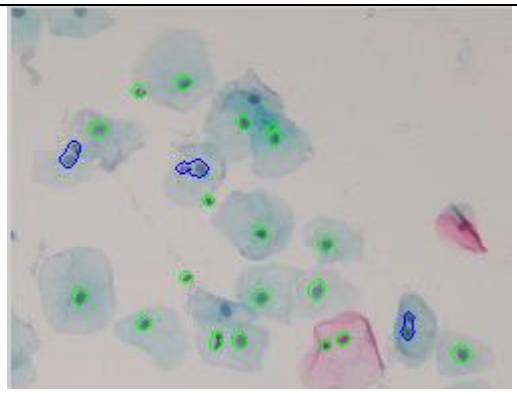
Figure 3.22. Feature space of adjacent and single cells with centroids from one of the sample images

3.6.4.3. Comparision of K-Means and Fuzzy C-Means

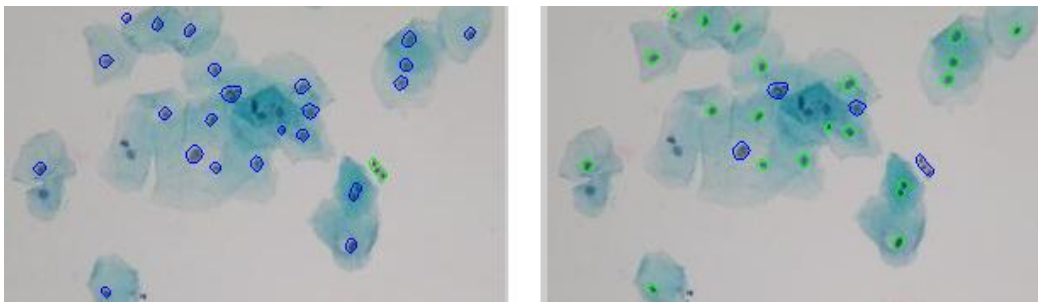
During the study, both of these clustering algorithms are examined with same samples and conditions. All key features extracted from the samples were same.

It is observed that final centroids of the clusters were very close with each of the clustering algorithms so that their success rate was similiar with the test group. Some of the test outputs are shown in Table 3.4. with elapsed time for both of the algorithms.

Table 3.4. Comparision of K-means and Fuzzy C-Means

K-Means	Fuzzy C-Means
Sample 1	
Elapsed Time: 0.1335 Seconds	Elapsed Time: 0.0180 Seconds
	
Sample 2	
Elapsed Time: 0.1066 Seconds	Elapsed Time: 0.0180 Seconds
	
Sample 3	
Elapsed Time: 0.1065 Seconds	Elapsed Time: 0.0143 Seconds
	

As it is seen on the Figure 3.24. results were similiar with both of the clustering methods but obviously k-means method needs much more time for completing the process also k-means works with lower success rates if number of the nuclei is decreased. One of the samples which FCM is more succesfull than k-means is shown in Figure 3.23.



(a) (b)

Figure 3.23. (a) A section from one of the samples which k-means miss some of the adjacent nucleus regions. (b) Performance of Fuzzy C-Means is better with fewer nucleus

Fuzzy C-means clustering method was chosen for eliminating adjacent cells based on the test results.

3.6.5. Finalization of Nuclei Segmentation

After the detection and elimination of adjacent cells, all detected nucleus were labelled. This final step aims to choose best nucleus for cytoplasm segmentation. Nucleus outside the detected cytoplasm regions or on the boundary of cytoplasm areas may cause loss of accuracy during the cytoplasm segmentation process. So that, all detected nucleus were classified into 5 main groups which is shown in table 3.5.

Table 3.5. Labeled regions after adjacent cell analyses

1	Nucleus inside the total cell area
2	Nucleus outside the total cell area
3	Nucleus on the total cell area boundary
4	Adjacent cell nucleus
5	Unidentified pixel group

Developed system marks the group 1 with red, group 2 with green, group 3 with blue, group 5 with yellow and detected adjacent nuclei groups were pink colored and framed. Yellow marked objects may be analysed again at the following stages of analyses or they may be marked for further operator analyse. All groups are shown on a samples in Figure 3.24.

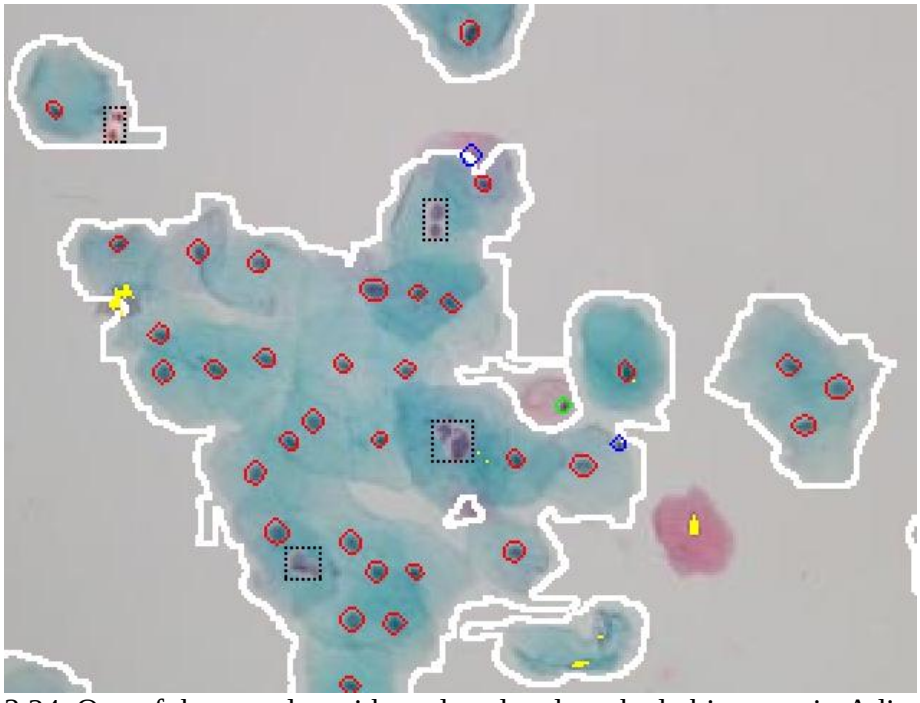


Figure 3.24. One of the samples with analysed and marked objects on it. Adjacent nucleus regions are pink, cytoplasm outher boundaries are white, nucleus boundaries are red and non-classified regions are yellow colored.

In addition to extracted features and gethered information about cell nuclei, developed system also collects several types of statistical data for following analyses

and possible operator check. An example of recent logged data until this stage is shown in table 3.6.

Table 3.6. Data logs after nucleus segmentation

Sample Size (Pixels)	480 x 640
Number of detected cell nuclei	97
Mean Mean Intensity of cell nucleus	128.3
Mean Max Intensity of cell nucleus	163.4
Mean Min Intensity of cell nucleus	78.2
Mean Area of cell nucleus	78.2
Mean Convex Area of cell nucleus	47.2
Mean Equivalent Diameter of cell nucleus	48.1
Mean Extent of cell nucleus	7.6
Mean Filled Area of cell nucleus	0.7
Time Elapsed (nucleus detection)	2.1 Sec.
Total cytoplasm Area (Pixels)	86647
Total cytoplasm Perimeter (Pixels)	15657
Division of Total area and total cytoplasm area	0.3
Time Elapsed (Background Extraction)	0.8 Sec.
Number of cell nuclei inside the total cytoplasm area	84
Number of cell nuclei outside the total cytoplasm area	6
Number of cell nuclei on the cytoplasm boundary	7
Number of adjacent cell nuclei	11
Time Elapsed (Nucleus Analyses)	0.2 Sec.
Number of detected artifacts	73

3.7. Segmentation of Cell Cytoplasm

Excessive growth of the cell nucleus is one of the most distinct symptom that gives clue about existence of cancer in the Pap-smear test samples. Consequently, relative size of the cell nucleus sould be one of the classification features for a computer aided Pap-smear screening system.

In the previous stages of this study, nuclei of the cells on the samples were detected and some key features were extracted for classification. Two of these features are area and diameter of the nucleus. However, in case of, all nucleus on the sample are large, it is not possible to use these features with clustering methods, or other non-linear classifiers. An example to case of most of the nucleus are abnormal is shown in Figure 3.25.

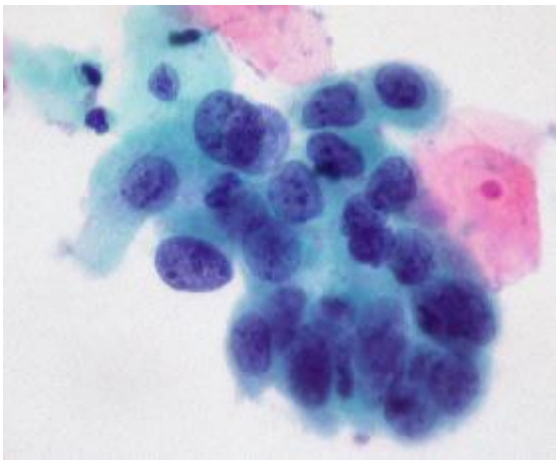


Figure 3.25. Almost all of the cell nucleus are abnormal in one of the sample images

Pre-trained machine learning algorithms may have ability to recognise these kind of size anbnormalities even if all cell nuclei on the sample are abnormal. On the other hand, machine learning algorithms only have ability to effectively recognize size abnormality due to nuclei area and perimeter if the sample is similiar with the samples in the training set. This is why, if an sample image is taken with different zoom and focus ratio, machine learning methods may misclassify objects if only feature about nuclei size is area and perimeter.

Thus, cytoplasm area of previously selected nuclei were segmented in this study. Features extracted from cytoplasm regions were used in addition to previously gathered nuclei data. It is possible to expect a healthy cell cytoplasm-nuclei area ratio inside a spesific margin. Even if sample is taken with a different zoom level, ratio shouldn' t be changed. Also cytoplasm area gives another key features about existence of malignancy such as cell density and shape.

Table 3.7. An example feature set whith cytoplasm data, used for classification by Jan Jantzen and George Dounias in 2006

COLUMN	FEATURE	NAME
B	Nucleus area	Narea
C	Cytoplasm area	Carea
D	N/C ratio	N/C
E	Nucleus brightness	Ncol
F	Cytoplasm brightness	Ccol
G	Nucleus shortest diameter	Nshort
H	Nucleus longest diameter	Nlong
I	Nucleus elongation	Nelong
J	Nucleus roundness	Nround
K	Cytoplasm shortest diameter	Cshort
L	Cytoplasm longest diameter	Clong
M	Cytoplasm elongation	Celong
N	Cytoplasm roundness	Cround
O	Nucleus perimeter	Nperim
P	Cytoplasm perimeter	Cperim
Q	Nucleus position	Npos
R	Maxima in nucleus	Nmax
S	Minima in nucleus	Nmin
T	Maxima in cytoplasm	Cmax
U	Minima in cytoplasm	Cmin

There are many factors that make it difficult to segment cytoplasm regions. First of all, samples are images of three dimensional enviroment which means samples give limited information about cells from one perspective. So that, most of the samples have overlapped and adjacent cytoplasm groups and regions. Even for a trained human eye, this kind of intricate regions may be hard to segment one by one. An example area to overlapped cell regions is shown in Figure 3.26.

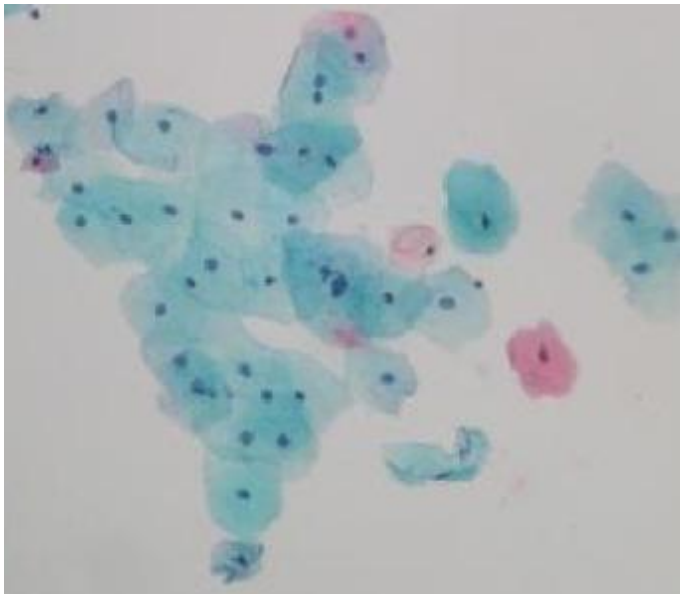


Figure 3.26. A part from a sample which have overlapped regions on it

In addition to mentioned difficulties, in most of the pap smear samples there are missing cell walls and also unwanted edges on cytoplasm regions which are not cell walls. A computer aided screening system should segment cytoplasm regions with an acceptable success ratio in spite of missing and unwanted information inside the cytoplasm areas.

In previous studies active contour model which is also known as snake model was frequently used for overcoming mentioned difficulties. According to that theorem analyse starts with an initial closed contour then it deformes itself until it takes the shape of the object which is desired to be delineated 8 (Kato, 2013)

Active contour model is an iterative method which based on minimisation of total energy with every new iteration. Total energy of the contour is sum of internal and external energies. External energy should be minimum on the maximum gradient magnitude points and internal energy should be minimum if the shape of the contour is regular and smooth (Tass, Witkin, Terzopoulos, 1988). Initial contour may expands like baloon or narrows like a rubber band for minimazing it' s energy. Total energy of the current contour is expressed with summation of three terms which is shown in (1)

$$E_{snake} = \int_0^1 E_{\text{int}}(v(s)) + E_{image}(v(s)) + E_{con}(v(s))ds \tag{3.13}$$

According to (3.13), E_{int} stands for energy due to bending of the contour. E_{image} stands for the gradient magnitude edges and E_{con} stands for the continuty of the points on the contour. Lowering these three components of the total energy causes the countour take the shape of the object iteration by iteration.

Active contour model is succesfull in many cases but this method is very senstive to distortion and easly effected by artifacts on the image. Besides, method is kind of an iterative process, which means, the algorithm may take long time to end the segmentation process (Williams, Shah, 1992).

In this study, a new deformable model is inspired from fluid dynamics for cytoplasm segmentation in Pap-smear image samples. Fluid mechanics allow us to simulate flooding of a fluid on the image. In the proposed method behaviour of the fluid is effected by forces on the image with every iteration. In the end of the operation, it was expected that, cytoplasm area of the selected cell should be defined by the total area of the simulated fluid.

One of the similiar approach which uses fluid flow is watershedding algorithm. In this method, segmentation is simulated by an immersion process. A topographic map of the image is used for flooding the image. Behaviour of the flood according to map defines the different regions in image (Qing Chen, Xiaoli Yang, Emil M. Petriu, 2004). Watershedding may be an effective and fast method for determining outlines of the adjacent cell groups but this method also may fail at the adjacent inner regions.

In the proposed method fluid behaviour is simulated as realistic as possible. Differently from watershedding method fluid can flow through some of the edges but others are behaves like walls to the fluid. In that way, it was possible to minimize the error caused by distortions. This selective behaviour is provided by some of the qualities of the fluids such as surface tension, diffusion and advection.

Navier-Stokes equations were used for modelling these qualities during the simulations. These equations basically depends on application of the newtons second law to fluid motion.

3.7.1. Fluid Dynamics Based Cytoplasm Segmentation

Gradient of the samples were converted into black and white half tone images by error diffusion method in the developed algorithm. In that way, sudden change regions were presented by large number of closer pixels and slow change regions were suppressed as few in number far pixels. All pixels, come from edge data of the image were modelled as single solid objects which were capable of interact with simulated fluid. So that fluid is effected by edge data and spread by every iteration.

Every previously detected nucleus area was assumed as initial density source of simulated fluid which behaves like a drop of water inside the cell. So that, fluid had capability of searching it' s own way between the pixels come from gradient and find the true boundaries. Closer pixels reacts with fluid as solids with more tightly bounded atoms. So that fluid diffuses harder. Also simulated fluid may bounce from pixels or pixel groups. So that, it is possible to flow through harder regions for simulated fluid which are behind the strong edges. Process steps of the proposed method is shown in Figure 3.27.

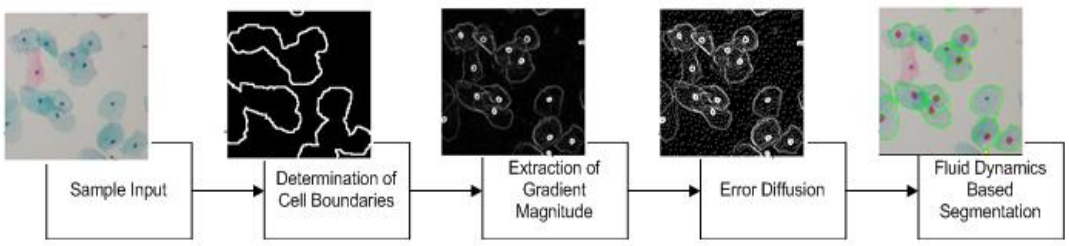


Figure 3.27. Basic block diagram of the proposed segmentation method

Some of the qualities of the fluids such as surface tension, diffusion rate, density etc. are optimised for the Pap-smear image samples. Thus modelled fluid have had ability to take shape of the cell with minimum influence of distortion.

Steps of the developed method are introduced below.

3.7.2. Error Diffusion

In the previous stages of the study a set of edge data is collected from sample images by morphological operations. This set includes cytoplasm and nucleus boundaries and also irrelevant edge data. Also, background data was extracted and some of the artifacts on the samples were eliminated.

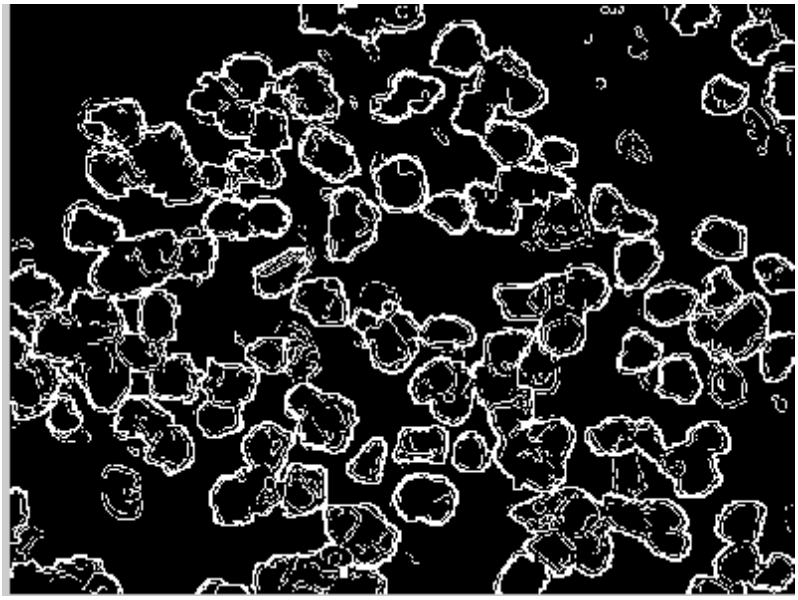


Figure 3.28. Total information gethered until this stage about cytoplasm boundaries

As it is seen on the Figure 3.28. there are many unwanted edge data inside the total information which is collected before. This unwanted data mostly comes from regions where cell overlapping occurs and also some of the cells may be folded on themselfs.

In the error diffusion stage, it was aimed to reconstruct edge data to allow the simulated fluid flow through edges which are not cytoplasm boundaries.

Error diffusion is one of the halftoning methods which depends on distrubution of quantization error to neighbourhood pixels. Main use of this method is converting multi level images into black and white images. After the process countinous edges of the images are represented by group of dots or pixels due to their intensity (Wu, 2005). A sample output is shown in Figure 3.29.

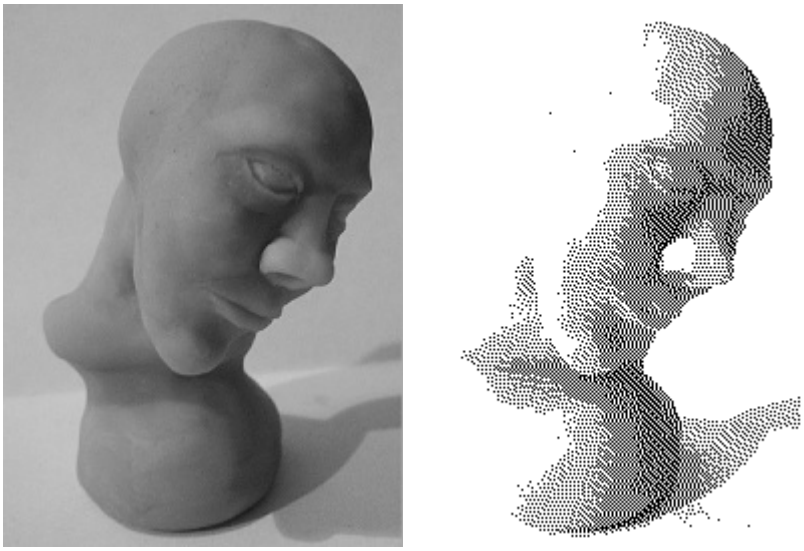


Figure 3.29. An example image before and after error diffusion process.

Floyd-Steinberg algorithm is used for dithering sample images by error diffusion in this study. According to that algorithm halftoning process based on re-quantization of every pixel of the image. If there is difference between new and old values of the pixel, this difference is assumed as error of the pixel (Omohundro, 1990).

In Floyd-Steinberg dithering algorithm distrubution of this error value is achieved by convolution of a kernel with the image which is shown is (3.14.)

$$\frac{1}{16}\begin{bmatrix} & * & 7 \\ 3 & 5 & 1 \end{bmatrix} \qquad (3.14.)$$

In (3.14) " *" indicates the current pixel and blank space is for pixel which is processed before and doesnt need to be processed again. A sample region from one of the regions after process is shown in Figure 3.30.

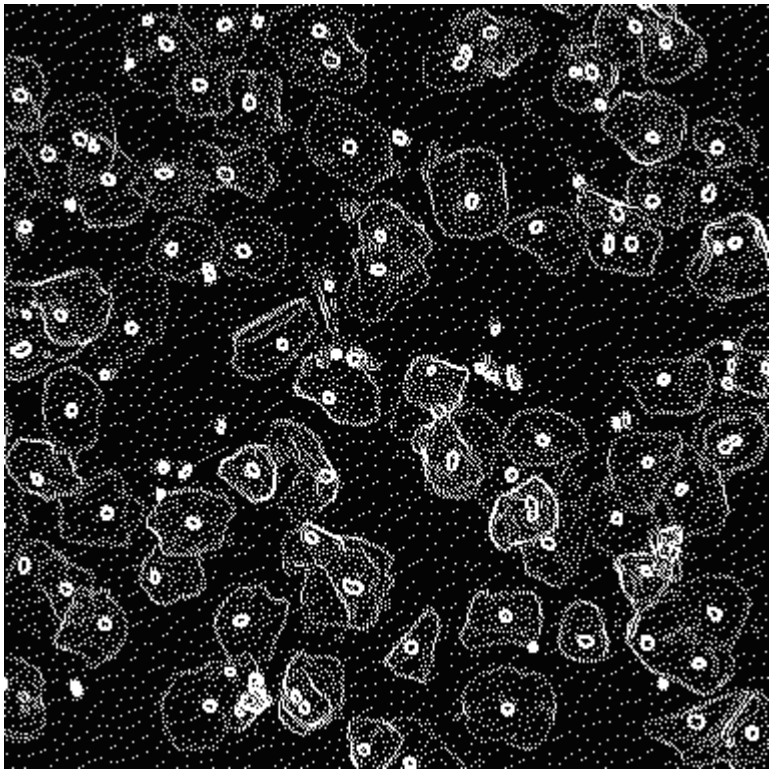


Figure 3.30. A part of one of the sample images after error diffusion process

Outher boundaries of the cells on the samples are determined in the previous stages so it is also possible to use these boundaries in addition to data retrieved after error diffusion process. These outer boundaries may help the prevent fluid spread out of the total cell area. All boundaries extracted until this stage is shown in Figure 3.31 on a sample.

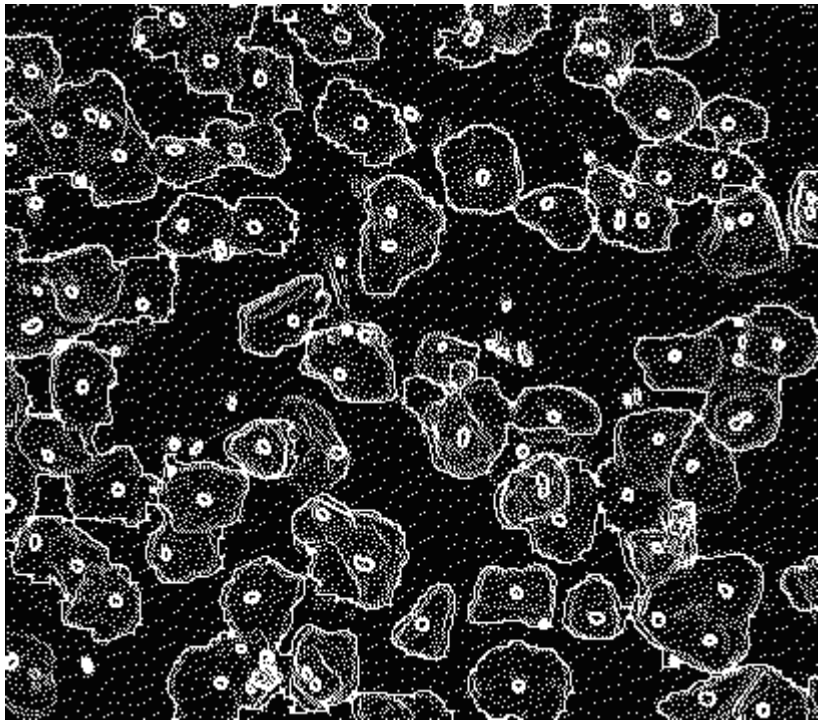


Figure 3.31. A part of a sample image after error diffusion with outher cell boundaries

3.7.3. Windowing

Developed algorithm tries to assign a cytoplasm region to every nucleus inside the total cell cytoplasm area. So that, process is started with cutting out of a window from image for every nucleus which may posibbly contains cytoplasm boundaries. One of the main reason of windowing is speeding up the iterations which needs to process all of the pixels in the active region.

Centroids of the cell nucleus are taken as center of the window which will cut out of the image. During the study size of the every processing window was 64x64 pixels. Which is large enough for searching entire boundary of the cytoplasm of selected nucleus.

Figure 3.32. Three of the windows taken from a sample image during the process by developed algorithm.

3.7.4. Fluid Dynamics

Simulating fluid behaviour on the image for defining shape and area is one of the most common segmentation methods in image processing such as watershedding algorithms. Unlike watershedding approach, method proposed in this paper is based on continious formulation which depends on Navier-stokes equations.

Mathematically, in the Eulerian (grid based) formulation, fluids are described by a velocity field u, a density field ρ and a pressure field p (Jalba, Roerdink, 2006). In this paper, simulated fluid was incompressible. Therefore fluid simulation was formulated with two components such as density and velocity vector fields . First one which is shown in (3.15.) is for formulating evolution of a velocity field over time.

$$\frac{\partial \mathbf{u}}{\partial t} = -(\mathbf{u} \cdot \nabla)\mathbf{u} + \nu \nabla^2 \mathbf{u} + \mathbf{f} \tag{3.15}$$

In equation (3.15.), term for pressure was removed (fluid which is desired to be simulated is incompressable) and f is denoted for the external forces which may apply to manipulate the flow of the fluid.

In equation (3.16.), density moving through the velocity field over time is formulate. Similiar to (3.15.), s is denoted for external density sources.

$$\frac{\partial \rho}{\partial t} = -(\mathbf{u} \cdot \nabla)\rho + \kappa \nabla^2 \rho + S \tag{3.16}$$

According to Navier-Stokes equations, every point occupied by fluid on the space has a velocity vector . Density of the fluid on that point changes over time due to that velocity vector.

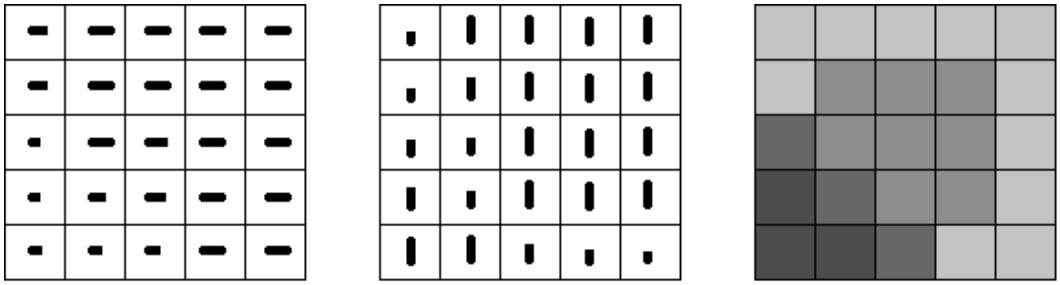
It is possible to simulate movement of the fluid by adding density and velocity sources on the image. It is also possible to react the simulated fluid with image by Navier-Stokes equations and calculate the change of a vector field in a time interval which belongs to simulating fluid.

There are a bunch of methods for numerical solution of the equations. In this paper, a fast and efficient computational grid method is preferred which is first introduced by Joe Stam in his " Real-Time Fluid Dynamics for Games" titled publication. Method, proposed in the publication is optimised for real time applications which means it works with high speed and requires less computational power.

Approach, developed by Joe Stam uses a computational grid which holds all the information about current state of an incompressable fluid. This grid consist of identical cells which are used for sample the fluid. Each cell updates with every iteration. In this study, each of the pixels assumed as an identical cell. Size of the main window used during this study is 64x64 which means simulated fluid is framed in a 64x64 sized window.

Every cell in the grid has a density and velocity value. Density value indicates the number of the particals in the cells. A matrix keeps the density value of the each grid cell during the simulation process.

Velocity is a vector quantity so that algorithm should hold it' s both magnitude and direction. There are two matrix for keeping velocity value of each grid cell which are visualised in Figure 3.33. First matrix represents magnitude in x direction and second represents magnitude in the y direction.



(a) (b) (c)

Figure 3.33. Representation of three matrix used for solution of Navier-stokes equations. (a) Matrix of the velocity magnitudes in the x direction. (b) Matrix of the velocity magnitudes in the y direction. (c) Matrix of the density values.

Particals in a cell may move by external forces or advection as an effect of conservation of mass. Also particals diffuse to neighbourhood cells (in this case pixels). Velocity vector field changes over time due to external forces and density changes. All of these effects causes particals move from one cell to other. Whereby developed algorithm have the general characteristics of the fluids. An example to interaction of each cell during diffusion is shown in Figure 3.34.

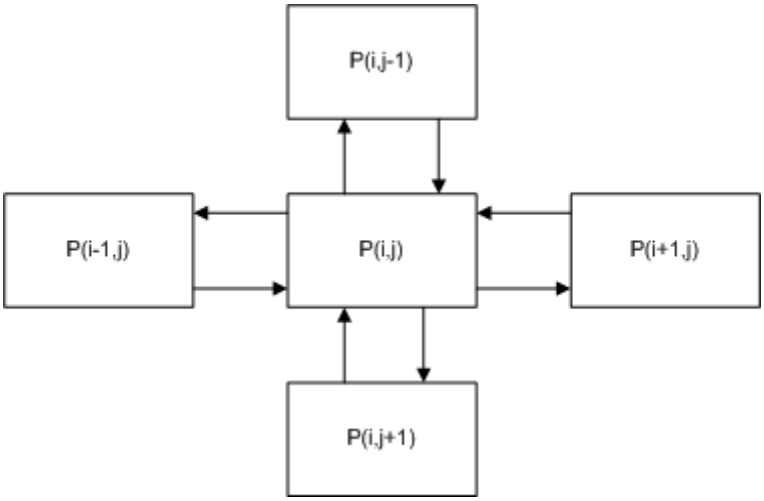


Figure 3.34. Each cell of the computational grid exchange particals via diffusion during simulation process

It was possible to simulate different fluids by changing viscosity and diffusion rate. In this study, both of the viscosity and diffusion rates are optimised

empirically for better experimental results. Higher diffusion rate may cause the fluid spread out of the cell walls and fluid may stuck between edges if viscosity is too low.

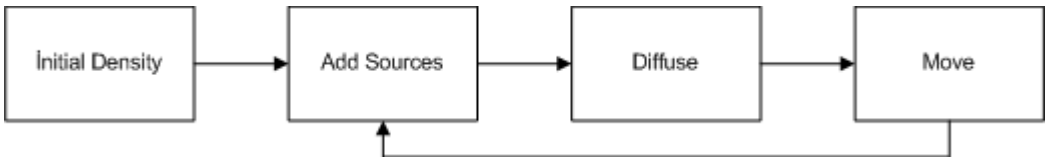


Figure 3.35. Block diagram of the movement process of the densities as proposed by Joe Stam in " Real-Time Fluid Dynamics for Games"

In first stage of the fluid dynamics based segmentation, inital density source is a fixed perimeter circle which is on the the exact centroid of the cell nuclei area. All pixels in the circle have an initial density value. Then For initialising flood simulation, eqaul external forces are applied to all directions. Then simulated fluid flew through particals which were obtained from error diffusion and react with them. Reaction is caused the fluid spread to cell walls. There were also pixels came from artifacts and distortions but mostly these pixels were not often enough to block the fluid. Whereby most of them were able to eliminated.

At the end of the simulation process, boundary of the simulated fluid was assumed as boundry of the cell which was desired to be segmented. Behaviour of the fluid in a sample window is visualise in Figure 3.36.

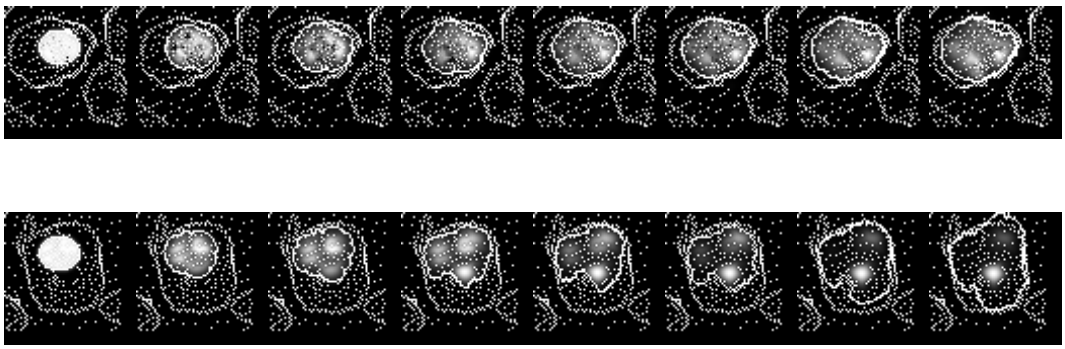


Figure 3.36. Simulation process with two different cells

3.7.4.1. Partical Interact

During the simulation process, movement orientation of the simulated fluid was determined by interaction of the fluid with particals which were obtained from error diffusion. This interaction causes the fluid having different densities at different areas. These differences helps simulated fluid to find it' s way through the partical groups. As a result of fluid behaviour, simulated fluid capable of pass through only some of the high density gradient regions. Which means decision mechanism of the proposed deformable model is based on interaction of fluid with particals. According to that approach, developed method completes one iteration in three main stages such as velocity step, density step and partical interaction step. Block diagram of simulation process is showed in Figure 3.37.

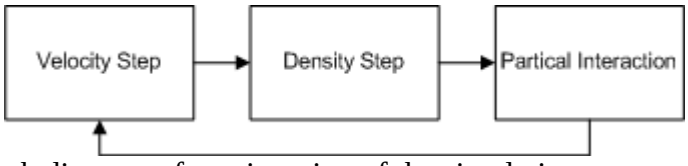


Figure 3.37. Block diagram of one iteration of the simulation process.

Unlike active contour model, even with adjacent or distorted areas, simulated fluid has ability to find it' s own way through particals. This capability is provided by selective behaviour of the particals on the image.

Selectivity is provided by permeability of the particals. Permeability of an interacted partical effects the density and velocity vector fields of the fluid.

If permeability rate is low at any region and particals are close enough, this region may block the fluid or rare particals with higher permeability rate may allow the some of the fluid pass through.

Because of time constraints, particals were only interacted with boundary pixels of the fluid and effect of a partical to boundary may change due to orientation of the fluid so that each of the boundary pixels were processed during all of the iterations for recalculation of reaction of the recent touching particals. Diffusion of simulated fluid through a pixel group is showed in three consecutive steps in Figure 3.38.

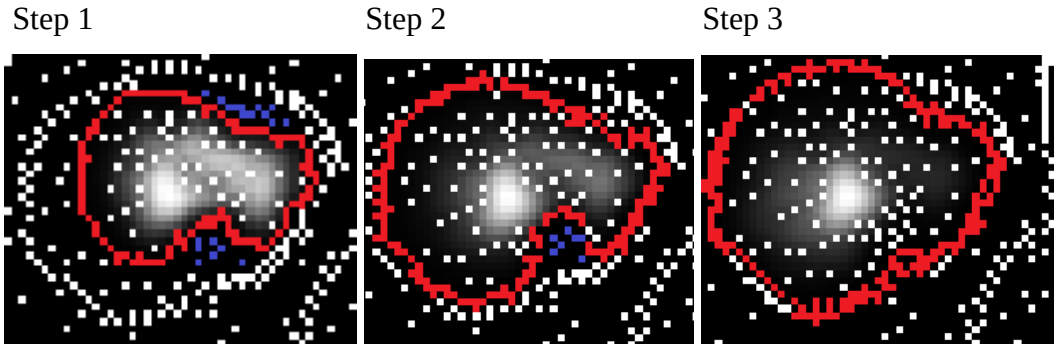


Figure 3.38. Boundary pixels of the fluid (red) may diffuse through some of the blocking partical (blue) areas

permeability rate of any partical, contacting to fluid boundary is determined by combination of three components. Which are shown in (3.17.).

$$P_{\text{Permeability}} = \alpha \cdot \frac{1}{P_{\text{Density}}} + \beta \cdot \frac{1}{P_{\text{Curvature}}} + \gamma \cdot \frac{1}{P_{\text{Line Score}}} \qquad (3.17.)$$

Each of these components are determined for an arc which is part of boundary of the simulated fluid. Arc has five pixels long and middle pixel is the recent processing pixel. A computational arc is shown in Figure 3.39.

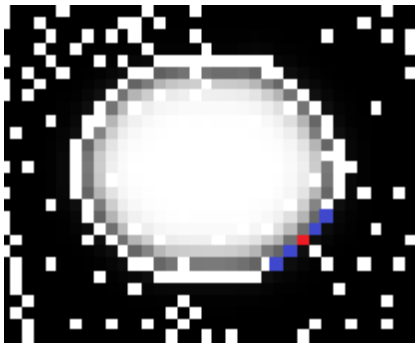


Figure 3.39. An example arc on a simulated fluid boundary.(Recent processing pixel is shown in red)

First term of the right hand side of the (3.17.) is partical density. It is basically total number of the particals in the neighbours of recent arc pixels. If numbers of the particals at the neighbours is high then diffusion rate will be

decreased. In case of α is increased fluid will spread faster but may diffuse out of the cell area.

Second term of the right hand side of the (3.17.) is curvature of the arc which is for preventing leaks. If there is a hole on a intense partical area fluid may tend to leak and spread away from an actual cell wall. So that if curvature of the recent arc is too high then permability rate of the particals, touching to the arc will be decreased.

If any edge of the gradient image is strong enough, it will represented by often and closer particals. Most of the distortions and edges which are not cell walls are also not that strong so mostly there were gaps between the pixels. If β is too low fluid may spread away from these gaps but if it is too high then fluid may have too strong surface tension which may possibly cause the fluid stuck. An arc with low permeability is shown in red in Figure 3.40.

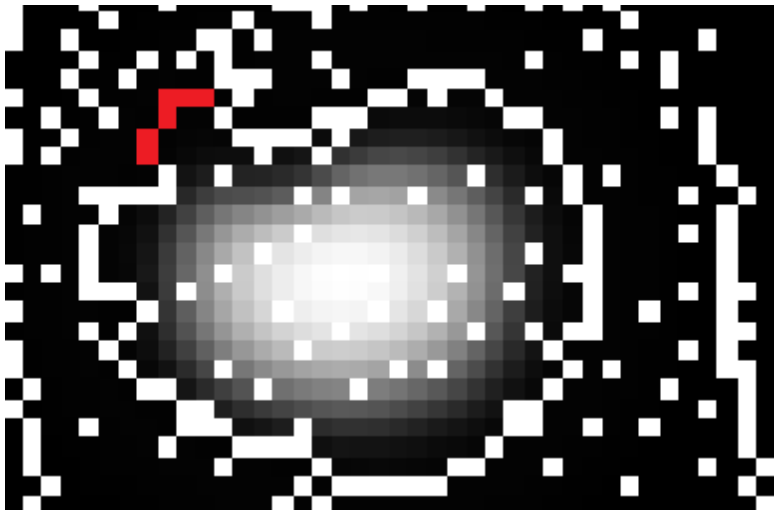


Figure 3.40. An arc (red) which has a strong curve and low permeability.

Third term of the right hand side of the (3.17.) is line score of the touching partical. Line score expresses the number of the pixels on the neighbourhoods of the touching pixel which are in line formation. Two pixel groups with different line scores are shown in Figure 3.41.

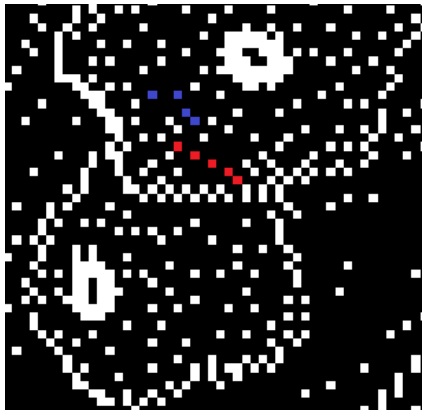


Figure 3.41. Line score of the red pixels is higher then the blue ones.

If line score of any partical is high then it' s possibility to belonging to an actual cell wall is increased. Bigger γ may cause the fluid doesn' t recognise cell-wall-like structures but also too low γ may cause fluid stuck between particles.

3.7.4.2. Stop Criteria

One of the parameters which directly effects on success of the proposed model is number of the iterations. Iteration count of each cell should be as low as possible because of time contrastraits.

Using same predefined count of iterations may decrease the succes of the model. Too many iterations may cause the fluid jump over the cell walls or simulation may stop before flooding to all possible areas if the number of the iterations is less then required. So that, a stop criteria for each cell is necessary for better performance. Result of over iteration on a sample cell is shown in Figure 3.42

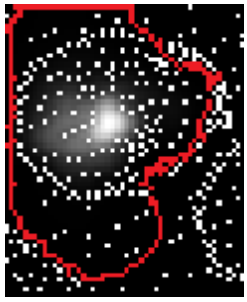


Figure 3.42. As a result of too many iterations fluid may diffuse through cell walls (Fluid boundary is shown in red).

In this study, growing acceleration of fluid area is selected as stop criteria. According to that approach, if fluid is flooded all the possible areas then there should be not much space to resume growing. If flooding is getting slower then process may end. So until propagation speed of the fluid is slowing down, simulation was continued. Some segmented test results are shown in Figure 3.43.

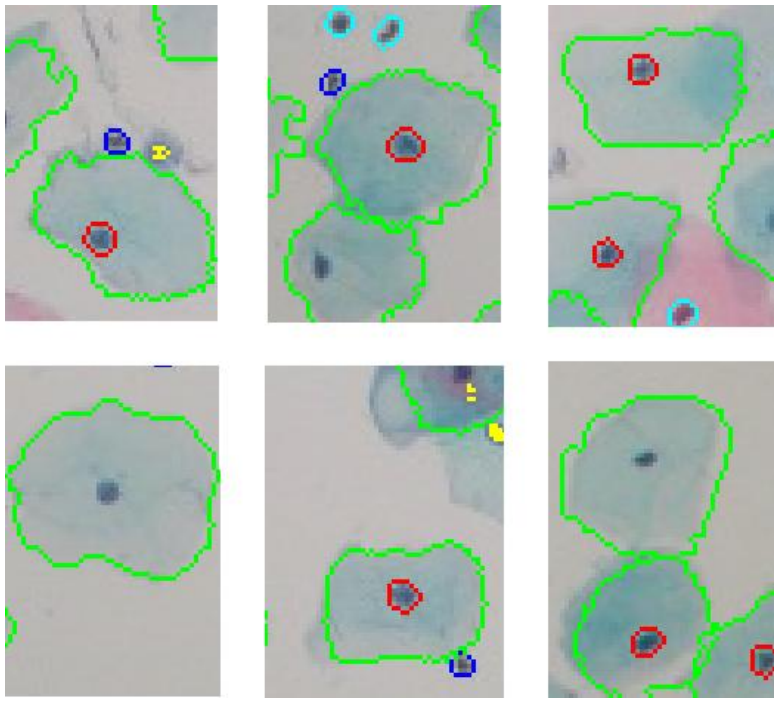


Figure 3.43. Some of the fluid boundaires at the end of the iterations.

3. 8. Ranking Cervical Cells

3.8.1. Feature Extraction

After segmentation process, Total area of cells were divided into individual cell regions which was consist two main parts such as cytoplasm and nucleus areas. Ranking of the cells was achieved due to features extracted from these two main divisions.

Radial basis function network was used for classification of the extracted features from segmented cell areas. Radial basis function networks require a training data set for learning the different classes of data which is desired to be ranked.

Herlev data set is used as training set for RBFN. There are total 917 cells divided into seven types of pre-classified single cells in the set. Firts three classes of set are normal type of cells and other four are abnormal ones which were ranked due to their abnormalities. These classified single cells were provided features of normal and abnormal cells to RBFN.

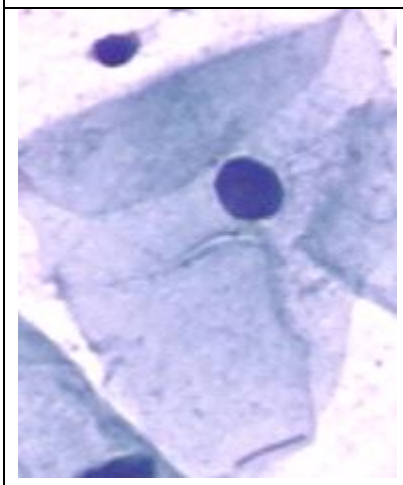
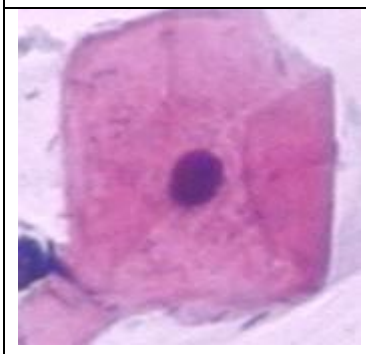
Table 3.8. Number of samples in seven classes of HDS used for training

Class NO.	Class Name	Number of Samples
1	Superficial Squamous	74
2	İntermediate Squamous	70
3	Columnar	93
4	Light Dysplasia	182
5	Moderate dysplasia	146
6	Severe dysplasia	192
7	Carcinoma in situ	150

All these classes listed on Table 3.8. have different characteristics which helps cytologs to diagnose malignancy on the cells.

12 features were examined in this thesis for classification of cervical cells which are listed for two sample images on Table 3.9.

Table 3.9. List of features extracted from two of the samples of HDS.

Picture of the sample	**Name of the feature**	**Value**
	Mean Intensity of nucleus	86,3
	Maximum Intensity of nucleus	123
	Minimum Intensity of nucleus	71
	Area of nucleus	799
	Convex Area of nucleus	830
	Equivalent Diameter of nucleus	31,8
	Extent of nucleus	0,7
	Filled Area of nucleus	799
	Eccentricity of nucleus	0,5
	Area of cytoplasm	27768
	N/C ratio	0,02
	cytoplasm mean intensity	192,9
	Mean Intensity of nucleus	82
	Maximum Intensity of nucleus	112
	Minimum Intensity of nucleus	68
	Area of nucleus	606
	Convex Area of nucleus	622
	Equivalent Diameter of nucleus	27,7
	Extent of nucleus	0,7
	Filled Area of nucleus	606
	Eccentricity of nucleus	0,6
	Area of cytoplasm	18043
	N/C ratio	0,03
	cytoplasm mean intensity	153,9

3.8.2. Classification

In this thesis, radial basis function networks were used as classifier for the extracted features.

A basic radial basis function network is two layered feed forward network which is very similiar to multi layer perceptrons. Output nodes of a RBF network is consist of linear summation functions as it is same in MLP' s (Bullinaria, 2004).

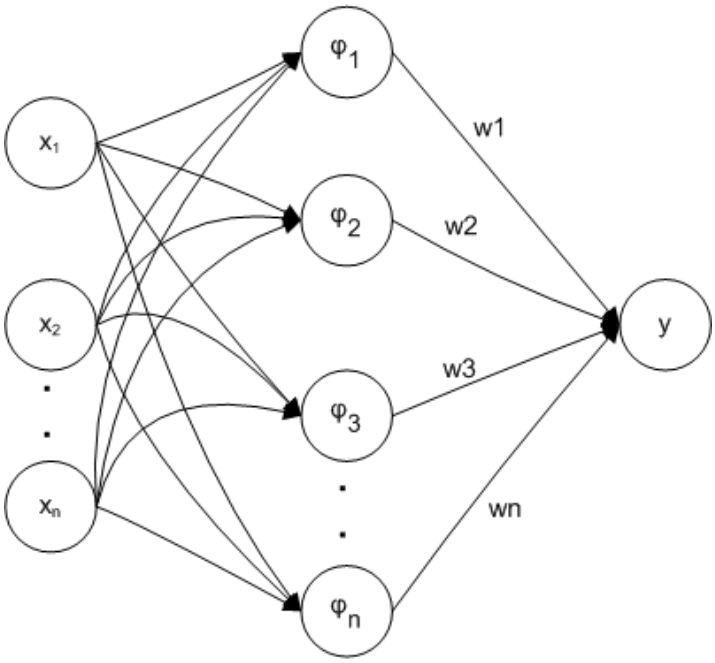


Figure 3.44. Basic structure of the radial basis function networks

In RBF networks there is a hidden layer, where neurons use radial basis functions as activation function. A function is classified as radial basis if it gives outputs that depend on the distance of the input from a given centroid. There are several types of radial basis functions such as multiquadric, inversemultiquadric and gaussian functions. Gaussian function which is shown in (3.18) were used as RBF in the network. In (3.18) "σ" indicates the spread of the radial basis function which determines the sensitivity. Larger spread may cause less sensitivity (Bullinaria, 2004).

$$\delta(i) = \exp\left(-\frac{i^2}{2\sigma^2}\right),\ i > 0 \qquad (3.18.)$$

Training of the network is achieved in two basic stages. First, weights between input and hidden layers than weights between hidden and output layers are determined. Training stage of the RBF networks is relatively faster then similiar networks.

3.8.2.1. Structure Of The Network Topology

Structure of the net, used in the study is consist of one hidden layer with RBF activation function, one input layer and one output layer. Topology of the constructed network is shown in Figure 3.45.

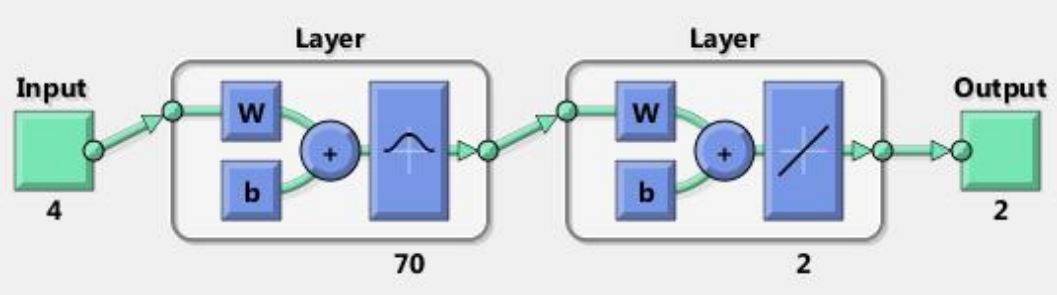


Figure 3.45. Topology of the network, used as classifier in study.

Output of the net is simply weighted summation of RBF outputs in the hidden layer, which is expressed in (3.19.)

$$y = \sum_{i=1}^{n} w_i \delta_i(\|x - t_i\|) \qquad (3.19.)$$

In (3.19.) " t" is used for centroids of the clusters. Which were change the output due to distance. If we substitute the distance between inputs and centroids in gaussian function we obtain the equation (3.20.) where "σ" stands for spread, " t" stands for centroids.

$$\delta_i = \exp\left(\frac{\|x - t_i\|^2}{2{\sigma_i}^2}\right) \tag{3.20.}$$

3.8.2.2. Training Of The Network

There are three set of values which are desired to be trained for classification in the net. Two of them are between input layer and hidden layer such as spread values and clusters. These two variables can be trained unsupervised. For example, centroids of the clusters are determined as in k-means clustering algorithm which is mentioned in previous sections of this study.

It is possible to determine spreads by normalisation method, expressed in (3.21.)

$$\sigma = \frac{\text{Maximum distance between two centers}}{\sqrt{\text{Number of centers}}} = \frac{d_{\max}}{\sqrt{m_1}} \tag{3.21.}$$

If we substitute (3.21) in (3.20) activation function of the hidden layer becomes (3.22).

$$\delta_i = \exp\left(-\frac{m_1}{{d_{\max}}^2}\|x - t_i\|^2\right) \tag{3.22}$$

Other set of values are weights between hidden and output layers. They can be determined by gradient descent method which is a supervised method. During the training process of the network, first centers and spreads should be determined then, while they are keeping fixed weights should be trained and updated. Block diagram of training process is shown in Figure 3.46.

Figure 3.46. Block diagram of the training process.

3.8.2.3. Performance Of The Network

It was observed during the experiments, there are three factors which may directly effect on performance of the network. One of these factors is quality of the samples in train set. In HDS there are many extreme samples for performing realistic tests and examinations but in this case for the sake of the efficiency, train set of the network was reduced to 70 clean and clear typical samples. Half of the choosen samples were normal and others were abnormal.

One of the other factors is used features for classification. If a feature shows small change between classes it may cause misclassifications. It is possible to analyse features and eliminate redundant features.

In this thesis, independent features significance testing method is used for feature elimination which is proposed by Weiss and Indurkhya (1997). It basically shows that which feature is most significant for the classification. According to that method significance degrees of extracted features are shown in Figure 3.47

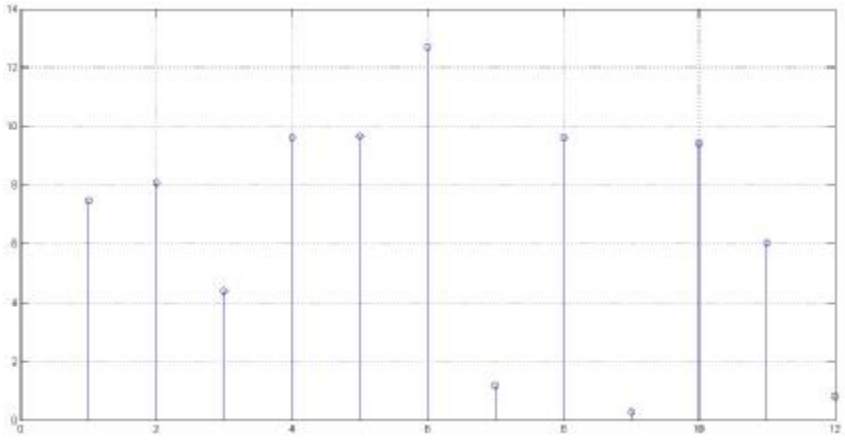


Figure 3.47. Significance degrees of extracted features.

Four features were choosen from most significant ones according to that measure. Choosen features are listed on Table 3.10

Table 3.10. Features, choosen for classification

1	**Area of nucleus**	Area of the nucleus as number of pixels.
2	**Equivalent diameter of nucleus**	Diameter of a circle with same area as nucleus as number of pixels.
3	**Extent of nucleus**	Ratio of total bounding box and area covered by nuclues in the total bounding box
4	**N/C ratio**	Ratio of nucleus and cytoplasm areas

Spread value of the RBF network is another factor which may directly effect performance. Different spread values were tested with 50 samples which were not included in the train set. Cells are classified as normal or abnormal. Efficiency of the network is measured with six different values which are accuracy, sensitivity, specificity, precision, recall, f-measure. Values are normalized to one which means one is best, zero is worst grade. Performance values of RBF network with different spread values is shown in Table 3.11.

Table 3.11 Performance table for different number of competitive neurons

Spread	accuracy	sensitivity	specificity	precision	recall	f-measure
10	0.62	0.62	0.62	0.62	0.62	0.62
50	0.86	0.86	0.86	0.86	0.86	0.86
100	0.92	0.92	0.92	0.92	0.92	0.92
200	0.90	0.90	0.90	0.90	0.90	0.90
400	0.94	0.94	0.94	0.94	0.94	0.94
800	0.98	0.98	0.98	0.98	0.98	0.98
1600	0.92	0.92	0.92	0.92	0.92	0.92

Best Performance is taken with spread 800 which is determined as spread value of the network

4.RESULTS AND SUGGESTIONS

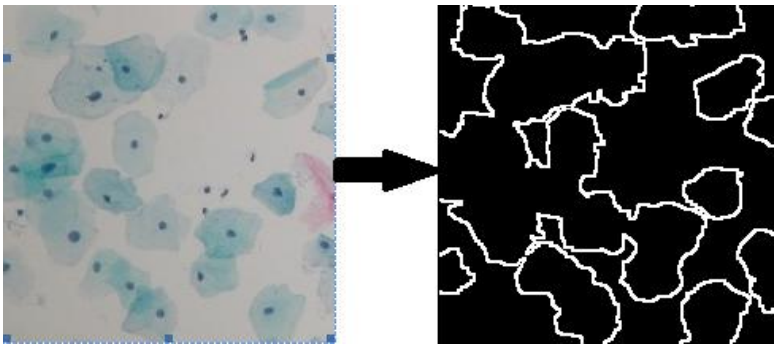
4.1. Background Extraction

Segmentation process is one of the most crucial parts of computer aided diagnosis systems. Many different approaches were examined in this study for getting better results with multiple samples.

During the tresholding stage of segmentation, fixed tresholds were used for every sample. In most of the cases selected HSV spectrum of pixels was efficiently cover cytoplasm areas.

It was seen that, fixed treshold method may seem efficient for limited sample sets but also it may cause serious loss of area with practical applications. Adaptive tresholding may solve the problem but operator decision may be the best solution. Outputs of two different treshold is shown in Figure 4.1.

(a)



(b)

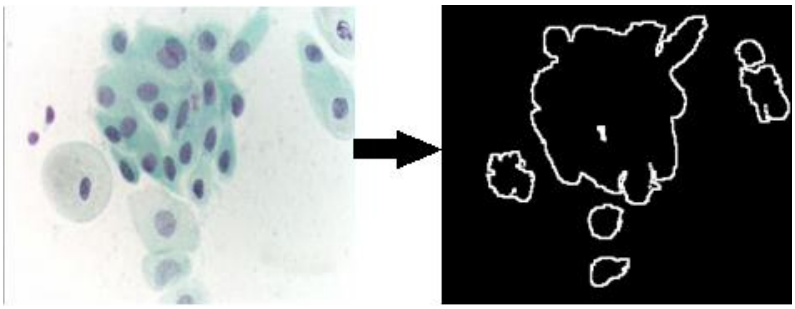


Figure 4.1. Two different region outputs for same treshold value

4.2. Nuclei Segmentation

Developed method works with high efficiency with most of the samples but scale of the samples has a great role on detecting nucleus boundaries succesfully. During this thesis study most of the samples were used in reduced scales and fixed zoom ratios. Bigger samples would decrease the segmentation success during practical applications.

Also developed algorithm sometimes looses some of the severe abnormal nucleus regions which may possibly cause misclassification. Checking outher neighbours of the boundary pixels of nucleus regions might help detecting missed nuclei parts. Some detected regions with missing pixel groups are shown in Figure 4.2.

(a) (b)

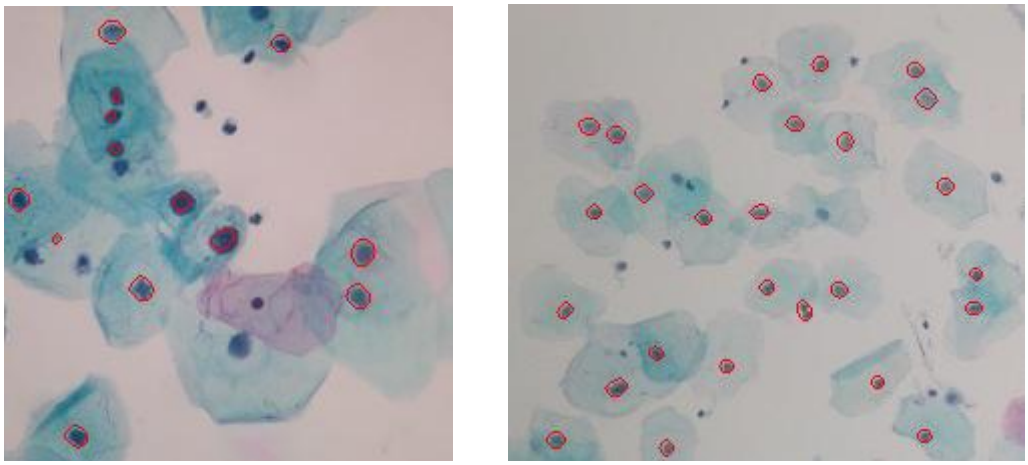


Figure 4.2. (a) A sample that detected nucleus boundaries with missed data on it. (b) One of the samples with succesfully detected nucleus boundaries

4.3. Cytoplasm Segmentation

In this thesis, cytoplasm segmentation task was achieved by a fluid dynamics based approach which have common points with both watershedding method and active contour model. Developed model is depends on flooding and interacting of a virtual fluid. An output of proposed model is shown in Figure 4.3.

This approach is capable of segmenting most of overlapped areas but also developed method might stuck in intense areas of the samples. Also in some cases, fluid leaks out of cell boundaries and causes wrong classification. Partical interaction rules may be optimized for leaks and some of the unnecessary particals, came from error diffusion stage might be removed by filtering.

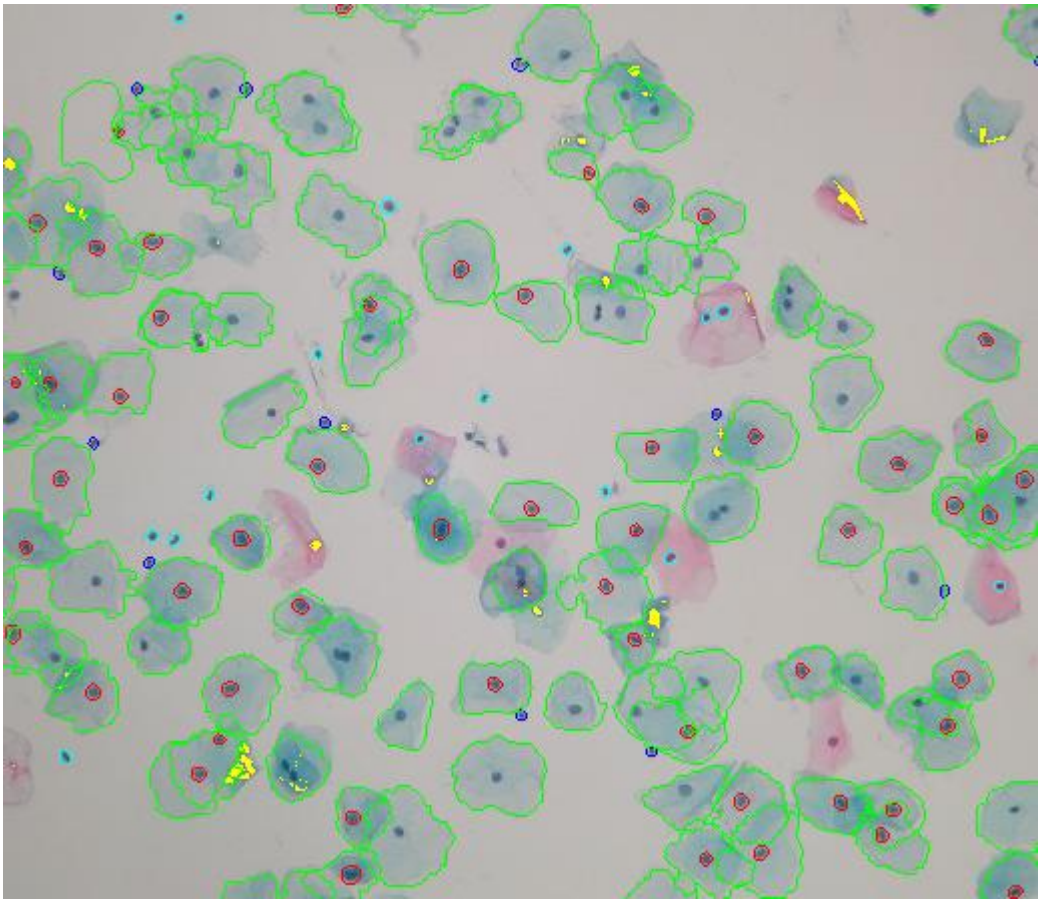


Figure 4.3. A sample with detected cytoplasm boundaries.

4.4. Classification

Radial basis function networks were prefered for classification of abnormal cells in this thesis. It was observed that, radial basis function network is highly efficient if it is working in predicted margins of data. All train samples were picked up from HDS during experimental process of the study and network was tested with typical samples with different sources. It was seen that, network is capable of

recognise abnormal cells but it is highly senstive to all type of differences even if cell is normal. A sample with classified cells on it is shown in Figure 4.4.

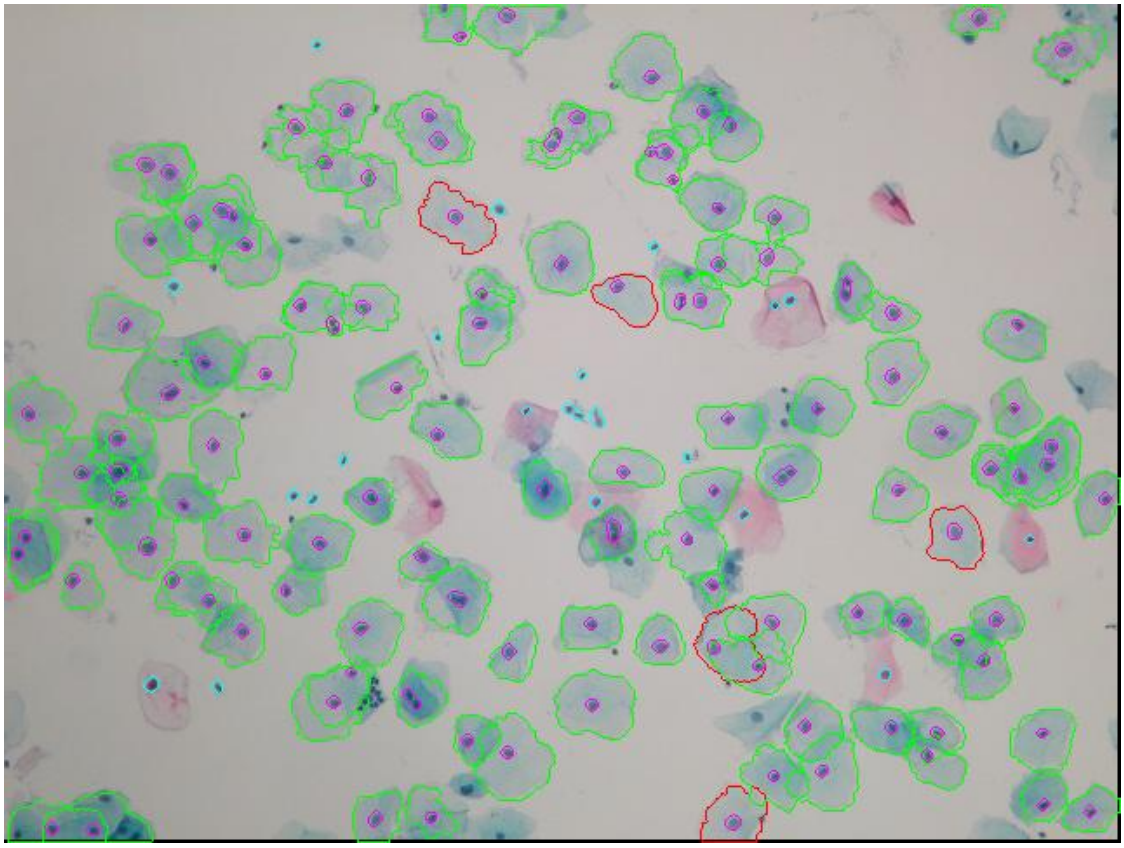


Figure 4.4. Classified samples which normal cells are shown in green and abnormal cells are shown in red.

It was also observed that, if there are missing nucleus pixels near the nucleus region, because of wrong measurement of nucleus area, network may not recognise the cell. A misclassified cell with missing pixels are shown in Figure 4.5.

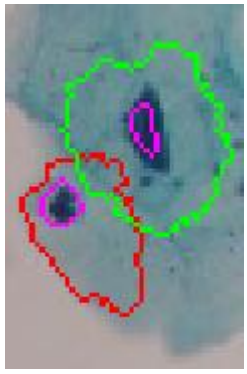


Figure 4.5 An example of misclassification because of undetected nucleus pixels.

If both of nucleus and cytoplasm regions are segmented with missing areas, network may classify the cell as normal. An example to mentioned misclassification is shown in Figure 4.6.

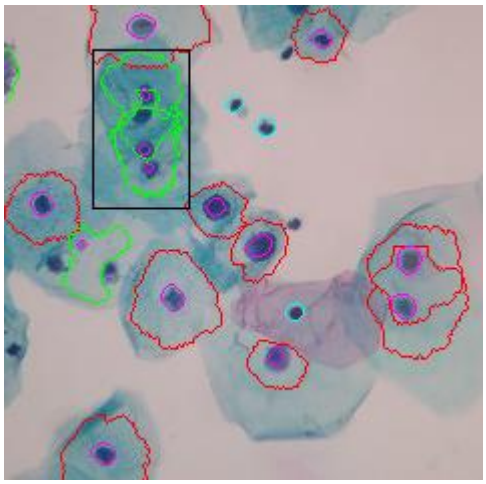


Figure 4.6. Wrong classified abnormal cells are shown in black box.

5.CONCLUSIONS

Development of a computer aided Pap-Smear diagnostic system was aimed in this thesis. One of the major expectations as a result from that study was a practical system which is capable of processing raw samples and giving numerical and visual data about existence of cancer on the samples.

Whole study was constructed in two main stages as segmentation and classification. Segmentation stage showed that conventional morphological operations may not be sufficient for a complete diagnostic system. There are both linear and non-linear problems to be solved for a efficient segmentation stage. Artifacts, intense and dark regions on the images, nucleus groups and overlapped cells causes morphological and linear operations fail. Low efficiency on segmentation stage may decrease the success of abnormality classification process.

Fluid mechanics were used at non-linear segmentation operation where it was not possible to apply a global rule for all of the samples such as cytoplasm area segmentation. Most of the pap-smear samples have overlapped cell areas and it was seen that these areas may cause misclassifications. Fluid behaviour was modeled during the study and a new fluid-image interaction method is applied for flooding the virtual fluid all over the cell which was desired to be segmented.

Fluid inspired segmentation was a victorious attempt for development of a better smart segmentation algorithm but there were also disadvantages of the approach.

Used method for solution of the Navier-Stokes equations (key of fluid behaviour) was depend on a computational grid which covers the entire window of process. Every cell of the grid should be processed again and again during the simulation for estimating fluid movement. This procedure spends a lot of time during the segmentation process. If there are too many nucleus on the sample, completing analyses may take three to ten minutes. Time consumption might be a great disadvantage for proposed method but it is also possible to optimise the source usage for faster computation. It was observed that proposed method would be faster with embedded hardware.

Radial basis function networks were used in this thesis for classification of the abnormalities. Clustering also might be useful for detection of abnormal regions but it was observed that clustering is not an efficient solution when most of the objects on the samples are abnormal.

It was also concluded that train set might be extremely crucial on network success. In case of extreme sample existence inside the train set, network would misclassify the input features. Also number of the samples in train set might be crucial. Smaller training sets may increase the possibility of misclassification and too big train sets might decrease the stability of the network.

Output of the network, constructed during the study, only informs that whether cell is abnormal. Output data of the proposed system doesn' t include degree of abnormality information. For a complete pap-smear diagnostic system, degree of abnormality would provide crucial data about level of malignancy.

It is possible to combine two different machine learning algorithms for a better classification. As an example to such a combination, detection of abnormal cells may achieve by learning vector quantization networks then a radial basis function network may score the abnormalitis on the cells. This two level classification method might provide much more data about malignancy degree of detected abnormal cells.

As a result of this study, it might be concluded that some of the common problems on computer aided cancer screening tests might be solved by nature inspired algorithms such as neural networks and fluid dynamics based segmentation methods.

REFERENCES

JALBA, A. C., Jos B.T.M. ROERDINK, J. B. T. M., 2006, A Physically-Motivated Deformable Model Based on Fluid Dynamics, 9th European Conference on Computer Vision, Graz, Austria

ANONYMOUS, 2003, CDC Publication #99-6949. Pg.1

BULLINARIA, J. A., 2004, Radial Basis Function Networks:Introduction

BURGER, G., JUTTING, U., RODENACKER, K., 1981, Changes in benign cell populations in cases of cervical cancer and it's precursors, 3(4), 261271

BYRIEL, J., 1999, Neuro-Fuzzy Classification of Cells in Cervical Cells

CENTERS FOR DISEASE CONTROL AND PREVENTION, 2013, http://www.cdc.gov/cancer/cervical/basic_info/screening.htm. (Access Date: 05/1/2013)

CENTERS FOR DISEASE CONTROL AND PREVENTION, 2013, Website of CDC http://www.cdc.gov/features/cervicalcancer/ (Access Date: 05/1/2013)

CHEN, Q., YANG, X., PETRIO, E. M., Watershed Segmentation for Binary Images with Different Distance Transforms Haptic, Audio and Visual Environments and Their Applications, 2004. Proceedings. The 3rd IEEE International Workshop on Date of Conference: 2-3 Oct. 2004

DOUGHERTY D. (Ed.), 2011, Medical Image Processing Techniques and Applications.Springer, New York, Page(s):11-237

HATHAWAY, R., BEZDEK, J., C., 2001, Fuzzy c-Means Clustering of Incomplete Data, IEEE IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 31, NO. 5

ISA, N. A. M., MASHOR, M. Y., OTHMAN, N. H., 2003, Contrast Enhancement Image Processing Technique On Segmented Pap Smear Cytology Images. In World Congress On Medical Physics And Biomedical Engineering (WC2003), 24-29

JANTZEN, j., DOUNIAS, G., 2006, Proceedings of the Nature-Inspired Smart Information Systems 2nd Annual Symposium. NISIS.

JACOBS, D., 2005, Image Gradients Class Notes for CMSC 426

KALE A., AKSOY S., ÖNDER S. , 2009, Cell Nuclei Segmentation in Pap Smear Test Images, Signal Processing and Communications Applications Conference, 2009. SIU 2009. IEEE 17th

KALRA, P., K., 2009, Canny Edge Detection
http://www.cse.iitd.ernet.in/~pkalra/
(Access Date: 05/1/2013)

KASS, M., WITKIN A., TERZOPOULOS, D., 1988, Snakes, active contour models, International journal of computer vision

KATO, Z., Snakes:Active Contours
http://www.cab.u-szeged.hu/~kato/variational/
(Access Date: 05/1/2013)

LI, Z., NAJARIAN, K., Neural Networks, 2001. Proceedings. IJCNN '01. International Joint Conference.

MACAULAY, C., PALCIC, B., 1990, An edge relocation segmentation algorithm. Anal. Quant. Cytol. Histol. 12(3), 165– 171

OMOHUNDRO, S. M., 1990, Floyd-Steinberg Dithering, Intemational Computer Science Institute

STEINBACH, T., GHOSH, K., 2006, ICDM: Top Ten Data Mining Algorithms K means

TANAKA, N., IKEDA, H., UENO, T., WATANABE, S., IMASATO, Y., 1977, Fundamental study of automatic cyto-screening for uterine cancer. II. Segmentation of cells and computer simulation. ActaCytol. 21(1), 79– 84

TANAKA, N., IKEDA, H., UENO, T., WATANABE, S., IMASATO, Y., 1977, Fundamental study of automatic cyto-screening for uterine cancer. I. Feature evaluation for the pattern recognition system. Acta Cytol. 21(1), 72– 78

TEKNOMO, K., K-Means Clustering Tutorials.
http:\\people.revoledu.com\kardi\ tutorial\kMean\
(Access Date: 05/1/2013)

WATANABE, S., 1974, the CYBEST group: An automated apparatus for cancer prescreening: CYBEST. Comput. Graph. Image Process. 3(4), 350– 358

WILLIAMS, D., J., SHAH, M., 1992, A Fast Algorithm for active contours and curvatur estimation CVGIP: Image Understanding archive Volume 55 Issue 1, Pages 14-26

WU, C., W., The Error Diffusion Halftoning Algorithm: Some Stability Results and Applications Beyond Halftoning

CURRICULUM VITAE

Çağlar CENGİZLER was born in Sivas at 02/08/1986. He has completed his high school education at Adana Ticaret Odası Anadolu Lisesi. He received his B.S. Degree in Electrical and Electronics Engineering Department of Atılım University in 2010. He has worked as research and development staff at several projects and companies. He works as lecturer in Adana Technical Sciences Vocational School recently.