Analysis of Cell Images for Identification of Cancer

a dissertation submitted in partial fulfilment of the
requirements for the *M. Tech. (Computer Science)*
degree of the Indian Statistical Institute

By

Parveen Gupta

Under the Supervision of

Dr. Deba Prasad Mandal

Machine Intelligence Unit

INDIAN STATISTICAL INSTITUTE
203, Barrackpore Trunk Road
Calcutta - 700 035

July, 1999

INDIAN STATISTICAL INSTITUTE

Telephone : (+91) (33) 577-8085/8086/2088
Extn. 3111
Telegram : STATISTICA, CALCUTTA 700 035
FAX : (+91) (33) 577-8092/8025
E-mail : dpmandal@isical.ac.in

203 BARRACKPORE TRUNK ROAD
CALCUTTA 700 035, INDIA

Residence :
A2/4, UTTARAYAN HOUSING ESTATE
192, B.T. ROAD, CALCUTTA 700 035
Telephone : (+91) (33) 556-9621

Dr. Deba Prasad Mandal
Lecturer
Machine Intelligence Unit

Dated: July 28, 1999

Certificate of Approval

This is to certify that the thesis titled *Analysis of Cell Images for Identification of Cancer* submitted by **Parveen Gupta** towards partial fulfillment of the requirements for the degree of *M. Tech. (Computer Science)* at the Indian Statistical Institute, Calcutta, embodies the work done under my supervision. His work is satisfactory. I wish him every success.

D. P. Mandal

Acknowledgements

At the outset, I express my gratitude to **Dr. D. P. Mandal** for his invaluable guidance, inovative suggestions and encouragement. At every stage of this work - from the initial planning of work up to the final preparation of report, he made serious efforts to improve the quality of the work. I am also thankful to his family, who might have been suffered when he worked tirelessly for long hours, for the final month.

In doing this project, a number of individuals have rendered their helpful suggestions, comments and encouragement; for them I would like to express my grateful thanks.

First, I am thankful to *Prof. S. K. Pal*, for providing me with resources, at MIU and introducing with Dr. D. P. Mandal. I am also thankful to *Dr. M. K. Kundu* for providing me with cell images. Special thanks are towards *Mr. B. Uma Shankar* for helping with many technical details. I would like to thank all staff members of Machine Intelligence Unit for supporting me during this work.

I wish to express my sincere thanks to *Prof. N. R. Pal* for his invaluable suggestions during this work.

I would like to thank members of Chitranjan National Cancer Institute, who have contributed diretly to this work. First, I would like to thank Dr. P. S. Basu, Dr. R. N. Chakrabarti and Dr. P. Das for providing me with the concept and knowledge about cancer and cell images. Thanks are due to Mr. S. S. Mandal, who helped me in all respect at CNCI. Finally , I would like to thanks all staff members of record section and histopathology department of CNCI, who helped me in many different ways.

I am also thankful to *Mr. P. S. Umesh* and *Mr. Pawan K. Singal* for discussion made with them.

Finally, I wish to pay my regards to my parents for their patience and devotion.

Parveen Gupta 21/07/93

Parveen Gupta

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cervical cancer or cancer of the cervix, a common kind of cancer in women, is a disease in which cancer (malignant) cells are found in the tissues of the cervix. It presents a significance challenge to the health care community as it is the second most common cause of cancer death among women world-wide [1, 2].

The present investigation is an attempt towards finding suitable methodologies to identify pre-cancerous changes in the tissues of cervix based on pattern recognition and image processing tools like thresholding, fuzzy c-means cluster seeking algorithm, mathematical morphology, textural feature etc.

1.1 Motivation

Cancer of the cervix usually grows slowly over a period of time. Before cancer cells are found on the cervix, the tissues of the cervix go through changes in which cells that are not normal begin to appear (known as *pre-cancerous cells*). Later, cancer cells start to grow and spread more deeply into the cervix and to surrounding areas.

Pre-cancerous and cancerous changes of the cervix can easily be found by *pap test* (also known as *pap smear*). This test involves scraping some cells from the surface of the cervix, smearing them on to glass slides and analyzing the slide images by looking the slides under a microscope.

Every slide image has two membranes (lines) known as *basal membrane* and *su-*

perficial membrane respectively. The portion between the basal and superficial membranes is only analyzed in the pap test. This portion has four different regions namely, i) *basal region* which lies just above the basal membrane, ii) *parabasal region* which lies above the basal region, iii) *intermediate region* which lies above the parabasal region and iv) *superficial region* which lies between the intermediate region and superficial membrane. Depending on availability of the immature/irregular cells in different regions, the images are characterized either as *normal* or *cervical intraepithelial neoplasias* (CINs) (i.e., CIN-1 (mild displasia), CIN-2 (moderate displasia) or CIN-3 (severe displasia)).

1.2 Scope of the Thesis

The present investigation is basically concerned with finding automatic methodologies to identify pre-cancerous changes in the tissues of the cervix as done in pap test. Initially a cooccurrence matrix (of gray values) based entropic thresholding method is applied on the slide images where mostly all the *cell nuclei* along with few other particles (lying in the stroma) are found to get segmented. The overlapping nuclei are then decomposed using a newly proposed algorithm. Each image is now rotated around its center (decided based on the nearly homogeneous region above the superficial membrane) to make superficial membrane (as well as basal membrane) horizontal and top side in the image. Using texture features *prominence* and *shading* on rectangular windows around every object pixel, the fuzzy c-means algorithm is found to extract out the homogeneous area above the superficial line and accordingly the superficial line is identified. Depending on the orientation of the (cell) nuclei, the basal line is found. The portion between basal and superficial lines are now decomposed into basal, parabasal, intermediate and superficial regions in the ratio 1:2:2:2 horizontally. The cell nuclei are analyzed to decide whether immature cells are found in different regions or not based on the number of cells, average size, maximum and minimum size, gray value variation, orientation etc. Accordingly the slide images are categorized as normal, CIN-1, CIN-2, CIN-3.

The rest of the report is organized as follows: In the second chapter, a brief introduction about cervical cancer is furnished. The third chapter provides a brief description of the existing pattern recognition and image processing tools which are

utilized in the present investigation. These include fuzzy c-means algorithm, thresholding, mathematical morphology, textural features etc. Our proposed methodologies for the analysis of slide images are described in the fourth chapter. In the fifth chapter, the effectiveness of the proposed methodologies on a few slide images along with a possible classification strategy is reported. The sixth chapter finds the conclusions. The report is ended with list of references.

Chapter 2

Cervical Cancer: A Brief Introduction

Cervical cancer or cancer of the cervix, a common kind of cancer in women, is a disease in which cancer (malignant) cells are found in the tissues of the cervix. The cervix is the opening of the uterus (womb). The uterus is the hallow, pear-shaped organ where a baby develops. The cervix connects the uterus to the vagina (birth canal). The part of the cervix closest to the body of the uterus is called the endocervox. The part next to the vagina is the ectocervix.

2.1 Signs and Symptoms

A cervical cancer patient is likely to experience any or some of the following symptoms, although each of these can be caused by problems other than cancer.

- Abnormal vaginal bleeding i.e., bleeding from the vagina after intercourse, between menstrual periods, before or after menopause, after coital or on straining.
- Watery, bloody discharge from the vagina; it may be heavy and smelly.
- Pain in lower abdomen or back regions.
- In later stages, a dull backache and general poor health.

2.2 Risk factors

In recent years, scientists have made much progress towards understanding the steps that take place in cells of the cervix during development of cancer. In addition, they have identified several risk factors that increase the odds that a woman might develop cervical cancer.

- *Age :* Similar to many cancers, the older a woman gets, the more likely her chances of developing cervical cancer. The peak years of incidence are between 45 and 55. The severity of dysplasia or abnormal changes in cells that are the first warning signs for cervical cancer, increases with age.

- *Marital History :* Early sexual activity has long been associated with increased rates of cervical cancer, because in teenagers, the cervix cells are still maturing and vulnerable. The delay in first pregnancy (or conception) and multiple pregnancies increase the risk of having cervical cancer. Having more than one sexual partners (or having a male sexual partner who has had multiple partners or a history of venereal disease) can put a woman at high risk.

- *Sexual transmitted viral infection :* Human Papilloma Virus (HPV) infection is the most important risk factor for cervical cancer. HIV (the AIDS virus) infection makes a woman's immune system less able to fight HPV and early cancers.

- *Smoking :* Smoking or passive smoking may contribute to the development of cancer.

- *Poor nutrition :* Poor nutrition increases risk, perhaps by depressing the immune system, so a woman is less able to fight HPV and cancers.

Although the above factors seem to be involved, not all of these will be present for every woman.

2.3 Pap Smear

Pre-cancerous and cancerous changes of the cervix can easily be found by *pap test* (also known as *pap smear*). Here, first of all, a speculum (a metal instrument) is inserted that keeps the vagina open so that the cervix can be seen clearly. Next, a sample of cells and mucus is lightly scraped from the ectocervix using a spatula. A small brush or a cotton tipped swab is used to sample the endocervix. These samples are then smeared on to glass slides. The slides are sent to the lab where specially trained technologists and physicians examine the samples under a microscope.

The most widely used system for describing Pap test results is the Bethesda System (TBS). The first category of TBS is within normal limits, which means that no signs of cancer or precancerous changes or other significant abnormalities were found. Precancerous cell abnormalities are classified as *mild*, *moderate*, or *severe dysplasia* or cervical intraepithelial neoplasia, CIN-1, CIN-2, or CIN-3.

Although the Pap test is considered by many doctors to be the most successful example of a test for cancer prevention and early detection, this test is not perfect. One of its limitations is that Pap smears are examined by humans, and that perfectly accurate analysis of the hundreds of thousands of cells in each sample is beyond the capability of the human eye and brain. Engineers, scientists and doctors are working together to improve this test.

Another approach to Pap test improvement is the use of computerized instruments that can recognize abnormal cells in Pap smears. The present investigation may be considered as an attempt in that direction.

2.4 Characteristics of slide Images [1-4]

Before introducing the basic characteristics of a slide image, let us now mention the general cellular morphology. Cells are essential units of the structures of all living organisms. Their structure varies with their function, but a large number of components are common to all. Every cell has two basic parts, namely nucleus and cytoplasm where the nucleus is surrounded by the cytoplasm.

In the present investigation, we are concerned with images of the tissues of a cervix.

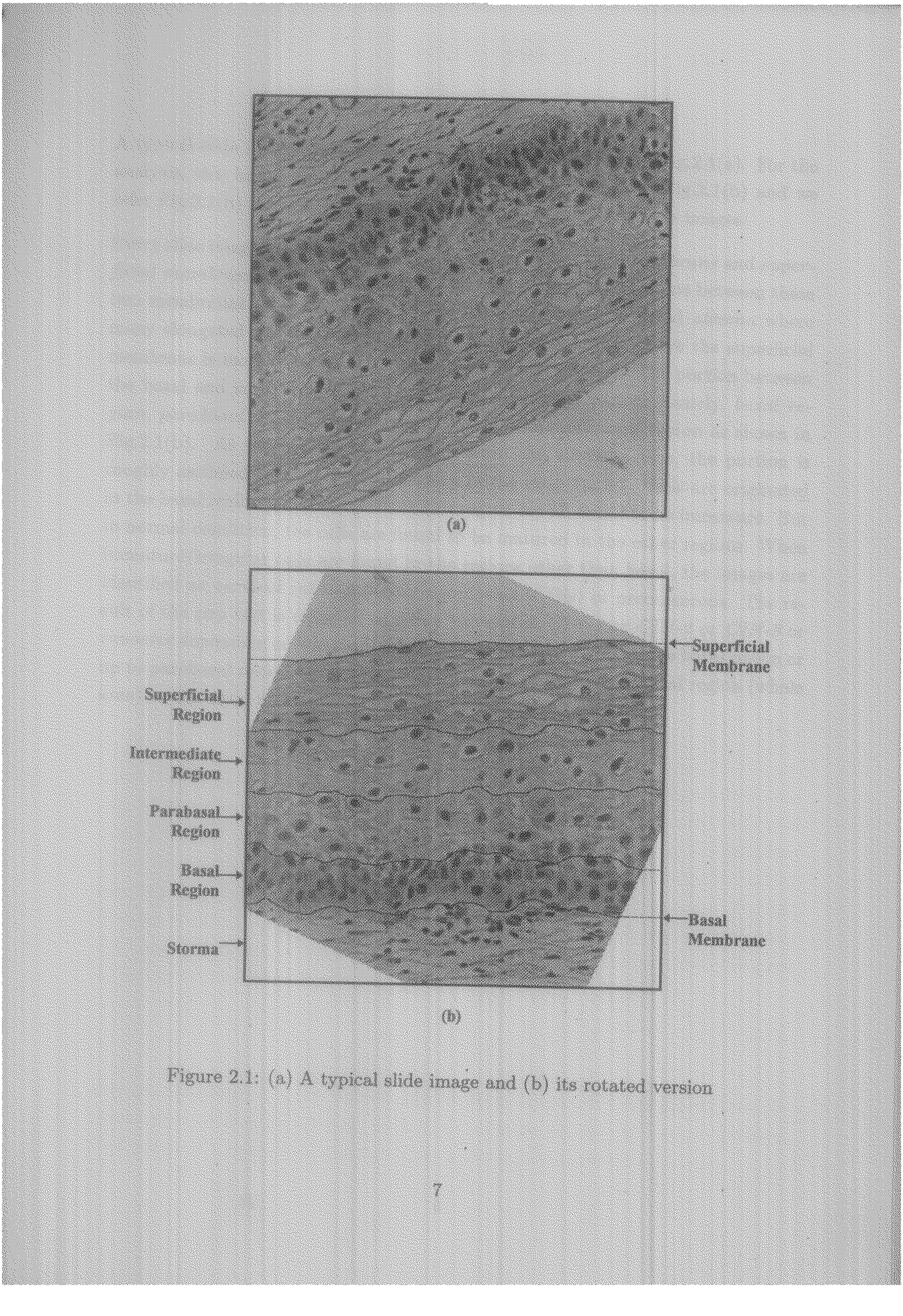


Figure 2.1: (a) A typical slide image and (b) its rotated version

A typical slide image observed under a microscope is shown in Fig.2.1(a). For the analysis, this image rotated to a particular angle as shown in Fig.2.1(b) and we refer Fig.2.1(b) for pointing out some basic characteristics of slide images.

Every slide image has two membranes (lines) known as *basal membrane* and *superficial membrane* respectively. Physicians analyze only the portions between these two membranes. The region below the basal membrane is called *stroma* where many elongated particles are found distributed. The region above the superficial membrane is more or less homogeneous with high brightness. The portion between the basal and superficial membranes has four different regions namely, *basal region*, *parabasal region*, *intermediate region* and *superficial region* as shown in Fig.2.1(b). As there is no boundary between the above regions, the portion is roughly assumed to have the four regions in the ratio 1:2:2:2. Cells are originated in the basal region and so the cells in this region are all found to be immature. But in normal condition, the cells are found to be matured in the other regions. When immature/irregular cells are found in the regions other than basal, the images are identified as *cervical intraepithelial neoplasias* (CINs) or precancerous. The result of the pap test is reported as either *normal* or *CIN-1* or *CIN-2* or *CIN-3* or *invasive* depending on whether immature cells are found only in the basal region or up to parabasal region, up to intermediate region or up to superficial region (whole area) or the basal membrane is found to be broken respectively.

Chapter 3

Pattern Recognition and Image Processing

In this chapter, we described some existing pattern recognition and image processing tools which are utilized in our present investigation.

3.1 Pattern Recognition

Pattern recognition and machine learning form a major area of research and development activity that encompasses the processing of pictorial and other non-numerical information obtained from interaction between science, technology and society. It can be viewed as a two-fold task, consisting of learning the invariant properties of a set of samples characterizing a class, and of deciding that a new sample is a possible member of the class by noting that it has properties common to those of the set of samples. The tasks required for developing and implementing the decision rule can be described as a transformation from the measurement space M to the feature space F and finally to the decision space D, i.e.,

$$\mathrm{M} \rightarrow \mathrm{F} \rightarrow \mathrm{D}.$$

Pattern recognition, by its nature, admits many approaches, sometimes complementary, sometimes competing, to the approximate solution of a given problem. The design of a pattern recognition system is a highly interactive process.

As we have discussed only the methods relevant to the present investigation, one can find details of various developments of pattern recognition in [5-10].

3.1.1 Clustering

Given any finite data set $\mathcal{X} \subset \Omega_X$ ($\mathbb{R}^N$) of objects, the problem of clustering in $\mathcal{X}$ is to assign object labels that identify natural subgroups in the set. Because the data are unlabeled, this problem is often called *unsupervised learning*, the word *learning* here referring to learning the correct labels for desirable subgroups. The objective is to partition $\mathcal{X}$ into a certain number (c) of natural and homogeneous subsets, where the elements of each set are as similar as possible to each other and at the same time, as different from those of the other sets as possible.

Many algorithms have been developed to obtain clusters from a given data set. Among those, the c-means algorithms and their generalization, the ISODATA clustering methods, are probably the most widely used. In the case of c-means algorithms, c is assumed to be known, whereas in the case of ISODATA algorithms, c is unknown. The performance of both models is influenced by the choice of c, the initial cluster centers, the order in which the samples are taken as input, the choice of distance measure and the geometric properties of the data.

Before describing the fuzzy c-means algorithm, a definition of fuzzy sets is provided.

Fuzzy sets [11]: A fuzzy set $\mathcal{A}$ in a space of points $\mathcal{X} = \{x\}$ is a class of events with a continuum of grades of membership and is characterized by a membership function $\mu_{\mathcal{A}}(x)$ which associates with each element in $\mathcal{X}$ a real number in the interval [0, 1] with the value of $\mu_{\mathcal{A}}(x)$ at x representing the grade of membership of x in $\mathcal{A}$. Formally, a fuzzy set $\mathcal{A}$ with its finite number of supports $x_1, x_2, \ldots, x_t$ is defined as a collection of ordered pairs

$$\mathcal{A} = \{(\mu_{\mathcal{A}}(x_i), x_i),\ i = 1, 2, \ldots, t\} \tag{3.1}$$

where the support of $\mathcal{A}$ is an ordinary subset of $\mathcal{X}$ and is defined as

$$S(\mathcal{A}) = \{x \mid x \in \mathcal{X} \text{ and } \mu_{\mathcal{A}}(x) > 0\}. \tag{3.2}$$

Assignment of membership function of a fuzzy subset is subjective in nature, and reflects the context in which the problem is viewed. It can not be assigned arbitrarily. •

Fuzzy c-means (FCM) algorithm [10,12]

Let $X = \{x_i \mid i = 1, 2, ..., t\}$ be a set of t vectors in N-dimensional feature space with coordinate labels $(x_{i1}, x_{i2}, ..., x_{iN})$ for x_i. Let $B = (\beta_1, \beta_2, ..., \beta_c)$ represent a c tuple of prototypes each of which characterizes one of the c clusters. The FCM algorithm minimizes the following objective function:

$$J(B, U, X) = \sum_{j=1}^{c} \sum_{i=1}^{t} (u_{ji})^m d^2(x_i, \beta_j) \tag{3.3}$$

subject to

$$u_{ji} \in [0, 1] \forall\ i, j$$

$$0 < \sum_{i=1}^{t} u_{ji} < t \quad for \quad all \quad i, j$$

and

$$\sum_{j=1}^{c} u_{ji} = 1$$

In equation (3.3), $d^2(x_i, \beta_j)$ represents the distance from feature vector x_i to the prototype β_j and $U = [u_{ji}]$ is a $c \times t$ matrix called the fuzzy c partition matrix where u_{ji} denotes the degree of membership of data point x_i in cluster β_j and m (>0) is called the fuzzifier.

From the necessary condition for minimizing J using the Lagrange multiplier technique we obtain the following update equations for membership values and centroids:

$$u_{ji} = \frac{1}{\sum_{k=1}^{c} [\frac{d^2(x_i, \beta_j)}{d^2(x_i, \beta_k)}]^{\frac{1}{m-1}}} \tag{3.4}$$

$$\beta_j = \frac{\sum_{k=1}^{N} (u_{ij})^m x_k}{\sum_{k=1}^{N} (u_{ij})^m} \tag{3.5}$$

The FCM algorithm alternates between equations (3.4) and (3.5) till either U or B stabilize. The algorithm can be initialized either on U or on B.

The problem with FCM is its high sensitivity to noise and the number of clusters is to be prespecified.

In the present investigation, the fuzzy c-means algorithm is used based on some texture features (which will be discussed in section 3.2.3) and assuming the value of m as 2.0.

3.1.2 Classification

The problem of classification is to assign every data point in the entire feature space to one of the possible (M) pattern classes. On the contrary, the clustering algorithms label the given data set $\mathcal{X} \subset \Omega_X$ ($\mathbb{R}^N$) into a number (c) of classes (i.e., subgroups). Classifiers are usually, but not always, designed with labeled data, in which case these problems are sometimes referred to as *supervised classification* (where the parameters of a classifier function $\mathcal{D}$ are learned). Many clustering algorithms are used as precursors to the design of a classifier when the only data available are unlabeled data. In such cases, the problems are sometimes referred to as *unsupervised classification.* In either case, the partitioning decision functions may be computationally explicit (*e.g.*, discriminant functions, nearest prototype rules) or implicit (*e.g.*, multilayered perceptron, k nearest neighbor rules) [5-9].

3.2 Image Processing

We have mentioned in the previous section that when the input to a pattern recognition system is a gray tone image, the measurement space usually involves processing tasks such as enhancement, filtering, noise reduction, segmentation, contour extraction, and skeleton extraction in order to derive salient features from the image pattern. This is what is generally known as *image processing.* These processed information are then used to compute various properties (*e.g.*, area, perimeter, centroid, *etc.*) and primitives (*e.g.*, line, corner, curve, *etc.*) of and relationships among the regions for developing decision rules/grammars. The ultimate aim is to enable the system to understand, recognize and interpret the input image pattern. Such a complete image recognition/interpretation system is called a *vision system* which may be viewed as consisting of three levels, namely low level, mid level and high level.

Basic principles and operations of image processing and computer vision are available in [13]-[16].

3.2.1 Thresholding

Thresholding is an important form of image segmentation where one wishes to identify and extract object regions from their background on the basis of differing brightness or gray levels. Many methods for the automatic selection of thresholds have been proposed [17-20]. Most of these methods base their selection on the optimization of some threshold dependent criterion function which is somehow related to the image and its properties.

One useful function is the entropy measure from information theory. Entropic thresholding can be done based on global image information (like histogram) or local image information (like co-occurrence matrix [18,19]). Some of these thresholding methods are guided by Shannon's entropy [21].

Histogram

Let $F = [f(x,\ y)]$ be an image of size $M \times N$, where $f(x,\ y)$ is the gray value at $(x,\ y)$, $f(x,\ y) \in G_L = \{0,\ 1,\ 2, \cdots, L-1\}$, the set of gray levels. A very crude summarization of the image details may be made through the gray level histogram, $h(g)$ = number of pixels in the image such that $f(x,y) = g$; g= 0, 1, ..., $L-1$.

Co-occurrence Matrix

The co-occurrence matrix of the image F is an $L \times L$ dimensional matrix that describes frequency of transition of intensity between adjacent pixels. In other words, the (i,j)th entry of the matrix gives the number of times the gray level 'j' follows the gray level 'i' in some specified ways.

Let 'a' denote the (i,j)th pixel in F and let 'b' be one of eight neighboring pixels of 'a', i.e., $b \in a_8 = \{(i,j-1),(i,j+1),(i+1,j),(i-1,j),(i-1,j-1),\ (i-1,j+1),(i+1,j-1),(i+1,j+1)\}$. Define $t_{gk} = \sum_{a \in F, b \in a_8} \delta$, where $\delta = 1$ if the gray level of 'a' is 'g' and that of 'b' is 'k', $\delta = 0$ otherwise. Obviously, t_{gk} gives the number of times the gray level 'k' follows gray level 'g' in any one of the eight directions. The matrix $T = [t_{gk}]_{L \times L}$ is, therefore, the co-occurrence matrix of the image F.

Different subsets of a_8 can be used to define different T [19]. Here, we use only

co-occurrence matrices computed by horizontal-right and vertical-down transitions i.e.,

$$t_{gk} = \sum_{i=1}^{M}\sum_{j=1}^{N}\delta \quad \text{with} \quad \delta = 1 \quad \text{if} \quad f(i,j) = g \quad \text{and} \quad f(i,j+1) = k$$
$$\text{or} \quad f(i,j) = g \quad \text{and} \quad f(i+1,j) = k$$
$$\delta = 0 \quad \text{otherwise.}$$

The entropy of the co-occurrence matrix gives another measure of image information. We call such entropy measure as local entropy or the *second order entropy* of the image.

Measures of Information

In the literature, most of the image segmentation algorithms that are developed using information theoretic concepts are primarily based on Shannon's entropy [21]. There are few methods which used the exponential entropy or quadratic entropy.

Consider a probabilistic system with n-states s_i; $i = 1, 2, ..., n$. Let p_i be the probability of the ith state s_i; $p_i \geq 0$, $i = 1, 2, \ldots, n$; $\sum_{i=1}^{n} p_i = 1$. Let $P_n = (p_1, p_2, ..., p_n)$.

Shannon's entropy [21]: For a discrete probability distribution P_n, Shannon, based on a set of axioms for a measure of information or uncertainty, derived the following unique definition of entropy :

$$H^1_{S_n}(P_n) = -\sum_{i=1}^{n} p_i \; log \; p_i. \tag{3.6}$$

It is a function of the form $H^1_{S_n} : \mathcal{P} \rightarrow [0, \infty)$ where $\mathcal{P}$ denotes the set of all probability distributions on finite sets. For all entropy definitions, *log* stands for log_2 This measure has the many interesting properties like expansibility, symmetry, continuity, monotonicity, maximum (for p_k=1 and p_i=0 for $i \neq k$), minimum (for $p_i = \frac{1}{n}$ $\forall i$), normalization, additivity etc. [22].

Thresholding using Cooccurrence Matrix

Let s be an assumed threshold. s partitions the image into two parts, object (pixels with gray values in $[0, s]$) and background (pixels with gray values in $[s+1, L-1]$).

Kapur et. al. [20] considered two probability distributions; one for the object and the other for the background. The entropy of the partitioned image is then defined by

$$H^1(s) = -\sum_{g=0}^{s}\left\{\frac{p(g)}{P_s}\ \log\ \frac{p(g)}{P_s}\right\} - \sum_{g=s+1}^{L-1}\left\{\frac{p(g)}{1-P_s}\ \log\ \frac{p(g)}{1-P_s}\right\}, \quad P_s = \sum_{i=0}^{s} p_i. \tag{3.7}$$

Kapur et. al. maximized $H^1(s)$ with respect to s to obtain the threshold.

Pal and Pal [19] proposed a few entropy based methods some of which maximize the second order local entropy of the object and background of the partitioned image. In other words, for an assumed threshold s, $0 < s < L-1$, the total second order entropy of the partitioned image,

$$H^2(s) = -\sum_{i=0}^{s}\sum_{j=0}^{s} p_{ij}\ \log\ p_{ij} - \sum_{i=s+1}^{L-1}\sum_{j=s+1}^{L-1} p_{ij}\ \log\ p_{ij}, \tag{3.8}$$

is maximized, where p_{ij} is the probability of occurrence of the pair (i,j) within the object/background.

The other method is based on conditional entropy of the partitioned image. Suppose an image has two distinct positions, the object O with gray levels $\{x_i\}$ and the background B with gray levels $\{y_j\}$. The conditional entropy of the object O given the background B can then be defined as

$$H(O/B) = -\sum_{x_i \in O}\sum_{y_j \in B} p(x_i/y_j) \times log\ p(x_i/y_j). \tag{3.9}$$

Similarly, the conditional entropy of the background given the object O is defined as

$$H(B/O) = -\sum_{y_j \in B}\sum_{x_i \in O} p(y_j/x_i) \times log\ p(y_j/x_i). \tag{3.10}$$

The pixel y_j, in general, can be the mth pixel after x_i. Since the estimation of such a probability is very difficult, we impose another constraint on equations (3.9) and (3.10) that x_i and y_j must be adjacent pixels. Thus,

$$H(O/B) = -\underbrace{\sum_{x_i \in O}\sum_{y_j \in B}}_{(x_i,y_j)\ adjacent} p(x_i/y_j) \times log\ p(x_i/y_j). \tag{3.11}$$

$$H(B/O) = -\underbrace{\sum_{y_j \in B}\sum_{x_i \in O}}_{(y_j,x_i)\ adjacent} p(y_j/x_i) \times log\ p(y_j/x_i). \tag{3.12}$$

The conditional entropy of the partitioned image can, therefore, be defined as

$$H^C(s) = (H(O/B) + H(B/O))/2. \tag{3.13}$$

The conditional entropy of the object given the background provides a measure of information about the object when we know about the existence of the background.

Algorithm *Cond_threshold*$(X,\ th)$

begin

Compute Co-occurrence matrix, T = $[t_{ij}]_{L\times L}$.
s = 0 ; **max** = 0 ;
th = 0 ; th is the threshold for segmentation
while $(s < L-1)$ do
compute $H_T{}^C$ by (3.13)
if ($H_T{}^C(s)$ > max) then **begin**
th = s ;
max = $H_T{}^C(s)$
end
s = s+1 ;
endwhile;
end;

3.2.2 Mathematical Morphology

Mathematical morphology [14, 16] is a tool for image processing which provides a system of operators. This system of operators when acting on complex shapes, decomposes them into their meaningful parts and filter out extraneous parts. They also serve to highlight spatial pattern and to remove spatial noise.

The primary morphological operations are dilation and erosion. Other morphological operations, such as opening and closing, are composed from dilation and erosion.

Let A and B be two sets in N-space (E^N) with elements a and b respectively, $a = (a_1, a_2, \ldots, a_N)$ and $b = (b_1, b_2, \ldots, b_N)$ being N-tuple of element coordinates. The dilation of $\mathcal{A}$ by $\mathcal{B}$ is denoted by $\mathcal{A} \bigoplus \mathcal{B}$ and is defined by

$$\mathcal{A} \bigoplus \mathcal{B} = \left\{ c \in E^N / c = a + b \quad for\ some\ a \in \mathcal{A}\ \ and\ \ b \in \mathcal{B} \right\} \tag{3.14}$$

Here $\mathcal{A}$ is associated with the image underlying morphologic processing and $\mathcal{B}$ is referred to as the structuring element, that shape which acts on $\mathcal{A}$ through the dilation operation to produce the result $\mathcal{A} \oplus \mathcal{B}$.

Erosion is the morphological dual to dilation. The erosion of $\mathcal{A}$ by $\mathcal{B}$ is denoted by $\mathcal{A} \ominus \mathcal{B}$ and is defined by

$$A \ominus B = \left\{ x \in E^N \;/\; x + b \in A \quad for\ every \quad b \in B \right\} \tag{3.15}$$

The dilation and erosion operations have been combined in a number of ways in our present investigation. The same structuring element is considered throughout this work which is of size 3×3, all with the object pixels and origin at the center i.e., in the position $(2, 2)$ as shown in Fig.3.1.

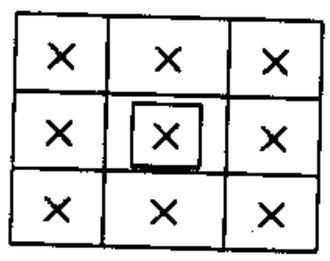


Figure 3.1: Structuring element

3.2.3 Textural Features [23]

Texture is often interpreted in the literature as a set of statistical measures of the spatial distribution of gray levels in the image. In the gray level co-occurrence matrix, the texture information in an image is constrained in the overall or average spatial relationships that gray level have with one another.

Let us consider a window of size n_1, n_2, i.e, with rows i_1, $i_1 + 1$, ..., $i_1 + n_1$ - 1 and columns j_1, $j_1 + 1$, ..., $j_1 + n_2$ - 1. For this window, we calculate the gray level co-occurrence matrix $[t_{gk}]$, $g,k = 0,1,\ldots, L$-1. The way of computing this matrix is described in section x.x.

Let $p_{gk} = \frac{t_{gk}}{R}$, where $R = \sum_{g=0}^{L-1} \sum_{k=0}^{L-1} t_{gk}$

From the co-occurrence matrix, the individual textural measures [23] are computed as described below,

- **Mean** (μ) gives the average value and it is taken as the arithmetic mean.

$$\mu = \frac{1}{N} \sum_{x=i_1}^{i_1+n_1-1} \sum_{y=j_1}^{j_1+n_2-1} f(x,y) \tag{3.16}$$

- **Standard Deviation** (s.d.) provides a measure of dispersion which is defined as

$$\sigma = [\frac{1}{N} \sum_{x=i_1}^{i_1+n_1-1} \sum_{y=j_1}^{j_1+n_2-1} (f(x,y) - \mu)^2]^{\frac{1}{2}} \tag{3.17}$$

- **Angular second moment** (A) gives a measure of homogeneity of the texture and is defined as,

$$A = \sum_{g=0}^{L-1} \sum_{k=0}^{L-1} (p_{gk})^2 \tag{3.18}$$

- **Contrast** (C) provides a measure of the local variation in the texture and is evaluated as,

$$C = \sum_{m=0}^{L-1} m^2 [\sum_{|g-k|=m}^{L-1} p_{gk}] \tag{3.19}$$

- **Entropy** (E) indicates the amount of randomness in the texture and is expressed as,

$$E = -\sum_{g} \sum_{k} p_{gk}\ log p_{gk} \tag{3.20}$$

- The measure **Homogeneity** (H) also provides an indication of the amount of homogeneity in the texture. It is computed as,

$$H = -\sum_{m=0}^{L-1} \frac{1}{1+m^2} \sum_{|g-k|=m}^{L-1} p_{gk} \tag{3.21}$$

- Cluster **Shading** (S) gives a measure of shading of the texture and is defined as,

$$S = \sum_{m=0}^{L-1} (m - \mu)^3 \sum_{|g-k|=m} p_{gk} \Big/ \sigma^3 \tag{3.22}$$

- The measure Cluster **Prominence** (P) also provides an indication of the amount of prominence in the texture. It is computed as,

$$S = \Big[\sum_{m=0}^{L-1} (m - \mu)^4 \sum_{|g-k|=m} p_{gk} \Big] \Big/ \sigma^4 - 3 \tag{3.23}$$

One can find many other texture features in the literature. As mention in section 3.1.1, these features have been used in FCM algorithm for segmentation.

Chapter 4

Slide Image Analysis

Our investigation for slide image analysis is described in this chapter. To provide an overall idea of the proposed system, an overview of the system is provided next.

4.1 Overview

The system takes an image as input which is analyzed and finally is classified either as normal or CIN-1 or CIN-2 or CIN-3. The block diagram of the system is furnished in Fig 4.1. It has seven blocks, namely, nucleus identification, rotation, basal membrane identification, superficial membrane identification, region identification, feature collection and classification.

The nucleus identification block mainly concerned with finding the nuclei of the cells present in the input slide image. In the next block i.e, rotation, the image is rotated to some extent such that the superficial membrane is placed horizontally at the top side. Note that the area above the superficial membrane is now more or less homogeneous with high gray values. This area is easily detected as a single cluster based on textural features shading and prominence. This is utilized in the superficial membrane identification block for marking the superficial membrane.

The basal membrane identification block then identifies the basal membrane using the fact that the (cell) nuclei above the membrane are vertically elongated but other object elements (not nuclei) lying below the membrane are horizontally elongated. The region identification block uses the marked basal and superficial lines for de-

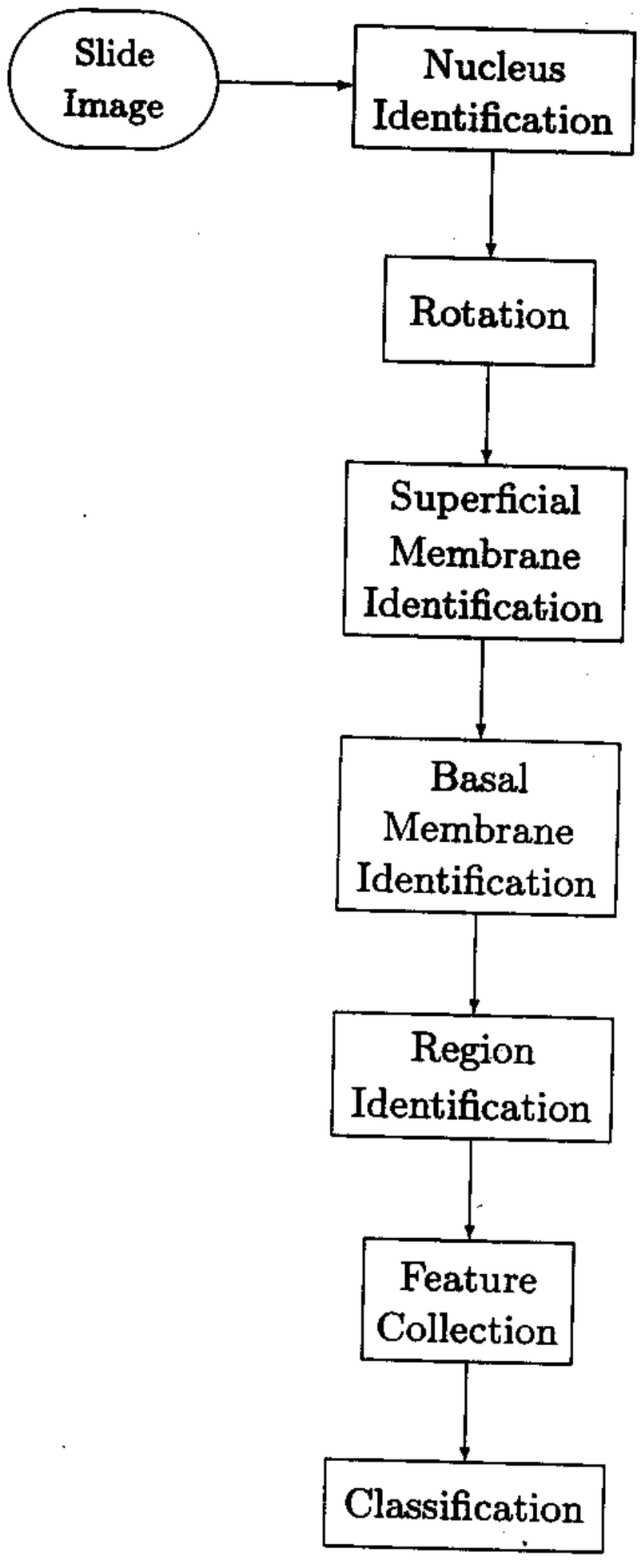


Figure 4.1: Block diagram

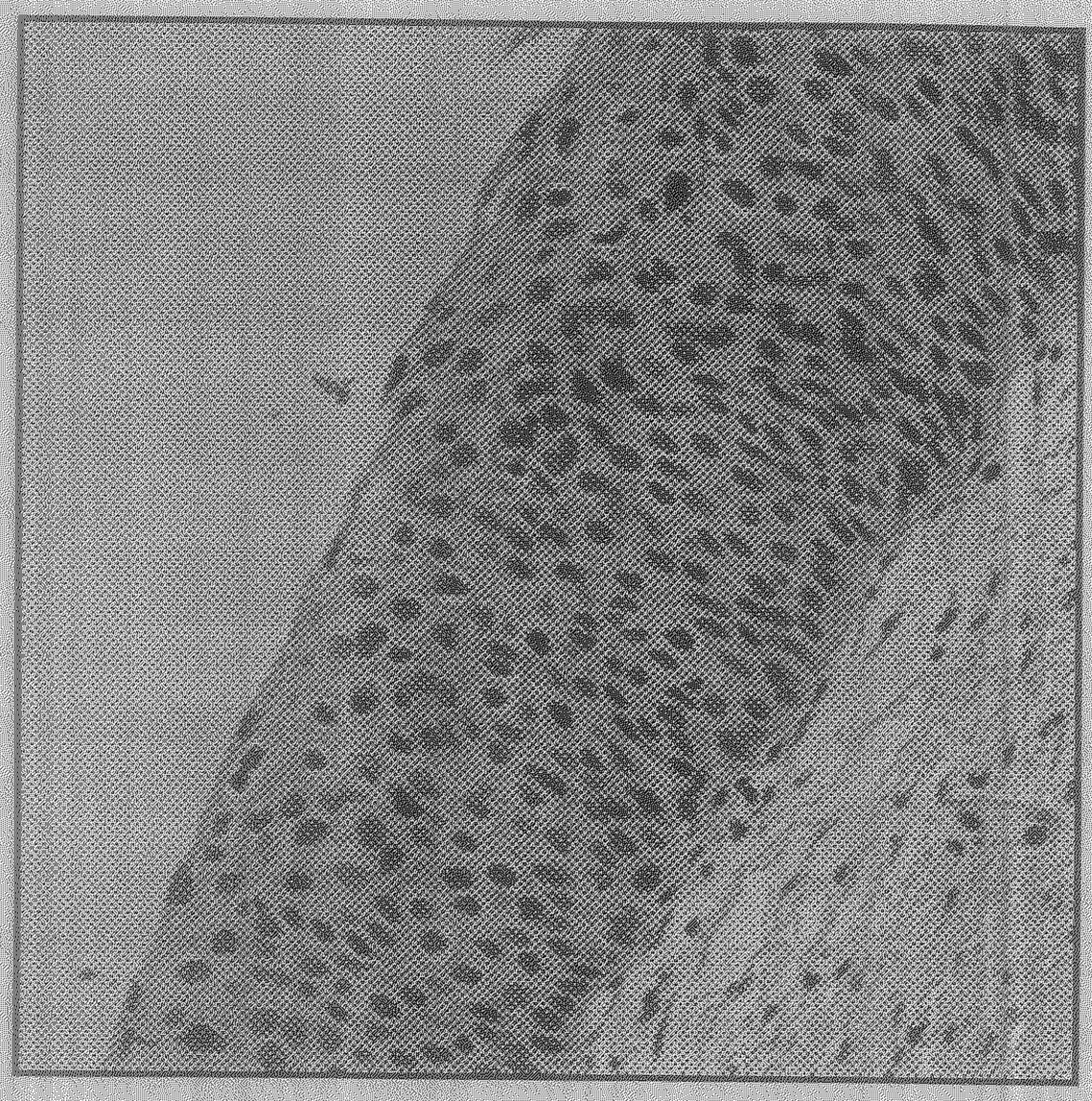


Figure 4.2: Reference Image

composing the area between these two lines into basal, parabasal, intermediate and superficial regions. Some characteristics features, namely, number of nuclei, nuclei sizes, gray value variation in nuclei, ratio of nuclei and non-nuclei pixels etc., are extracted for each of the four regions. Based on such features, a slide image may be classified as normal, CIN-1, CIN-2 or CIN-3.

Details of the first five blocks are presented in the following sections. we have described the last two blocks i.e., the feature collection and the classification blocks in the next chapter i.e., chapter 5. For a better understanding of our methodologies step-wise, we consider the slide image shown in Fig.4.2 throughout this section. After the description of a sub-process, it's effectiveness on this image is displayed.

4.2 Nucleus Identification

This section takes the original slide image as the input and identifies all the cell nuclei. It has three sub-process, namely, thresholding, local thresholding and morphological segmentation. The purpose of thresholding is to segment slide image into two parts as object and background regions such that the object region includes

most of the nuclei pixels along with some components of the storma region. The description of each nucleus is improved by applying the thresholding algorithm locally on an rectangular window around the nucleus. Finally a morphological based nucleus splitting algorithm is used to divide overlapped nuclei.

4.2.1 Thresholding

The histogram of the reference image [Fig.4.2] is furnished in Fig.4.3. Most of the histogram based valley seeking thresholding algorithm will find the threshold value around 200. Figures 4.4 (a), (b), (c) and (d) depicts thresholded image for threshold values 100, 120, 140 and 160, respectively. From these figures, it is evident that the most acceptable result for our purpose is obtained for threshold around 120. Similar is the case for all other slide images we have. Therefore, we have used here co-occurrence matrix based entropic thresholding method as described in section 3.2.1. This algorithm is able to extract majority of the nuclei pixels as object. The threshold obtained by this method on the image in Fig.4.2 is 117 and correspondingly segmented image is shown in Fig.4.5. Comparing this image with its original one, we can see some spurious (noisy) components also appeared as object pixels and these are removed by using morphological erosion operation (as described in section 3.2.2). As mentioned earlier, we consider the structuring element which is of size 3×3, all with object pixels and origin at the center i.e, in the position $(2, 2)$ as shown in Fig.3.1.

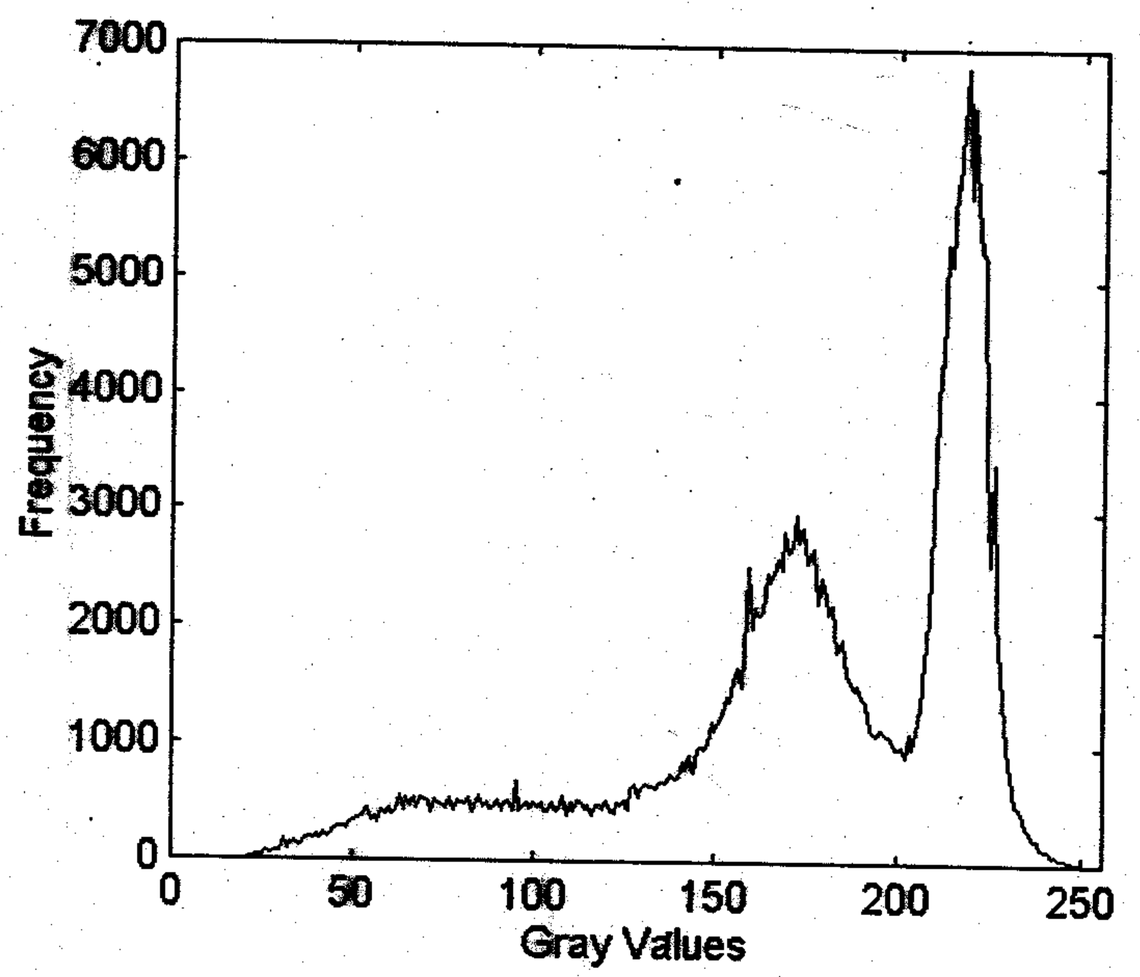


Figure 4.3: Histogram of the reference image

As a result of erosion operation, some spurious object elements get deleted. To get back the previous structure of other object components, a dilation operation is applied with the same structuring element. The output of Fig.4.5 after the above mentioned noise removal operation is shown in Fig.4.6. Each set of connected (8-neighbor) pixels are now onward referred to as one object item.

4.2.2 Local thresholding

Comparing the original image and the segmented image (obtained after the global thresholding), we can see that some of the object items are not exposed fully. Again a few of them are over exposed. So, there is a scope of improving these object items (nuclei). For this purpose, we applied the same thresholding algorithm locally on a rectangular window around each object item. We call this process as local thresholding. Depending on the size of object item, the size of window varies. The size of the rectangular window is extended by 5 pixels in all four sides of the smallest rectangle enclosing an object item.

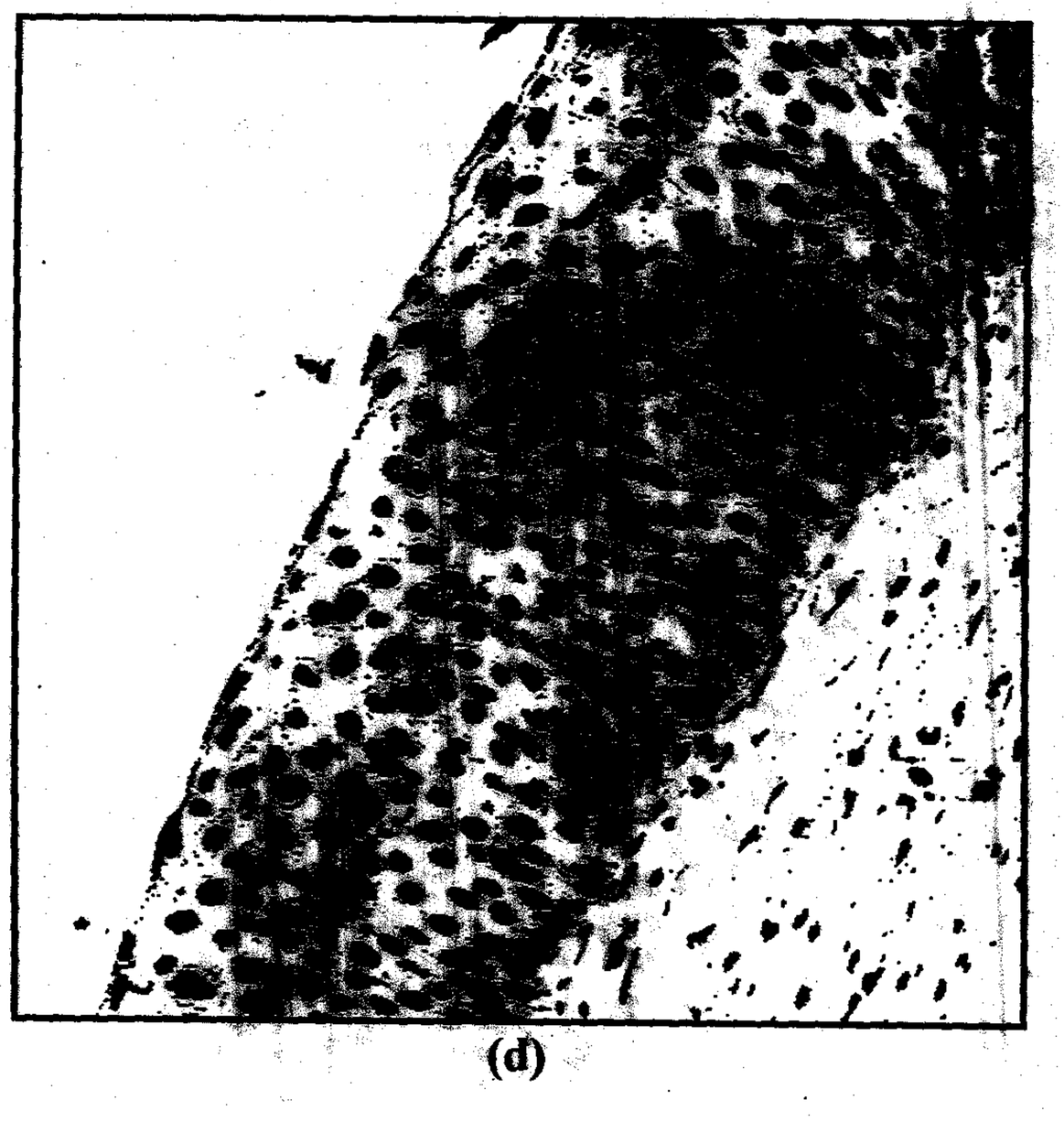
Only pixels which are classified as object, both globally and locally, are now called object pixels. The dilation operation is now applied thrice on the object pixels conditionally in such a way that the pixels are added under dilation only if they were labeled as object pixels by the local thresholding. The output using local

(a)

(b)

Figure 4.4: Thresholded reference image with thresholds as (a) 100, (b) 120, (c) 140 and (d) 160 .

(c)



(d)

Figure 4.4: (Continued)

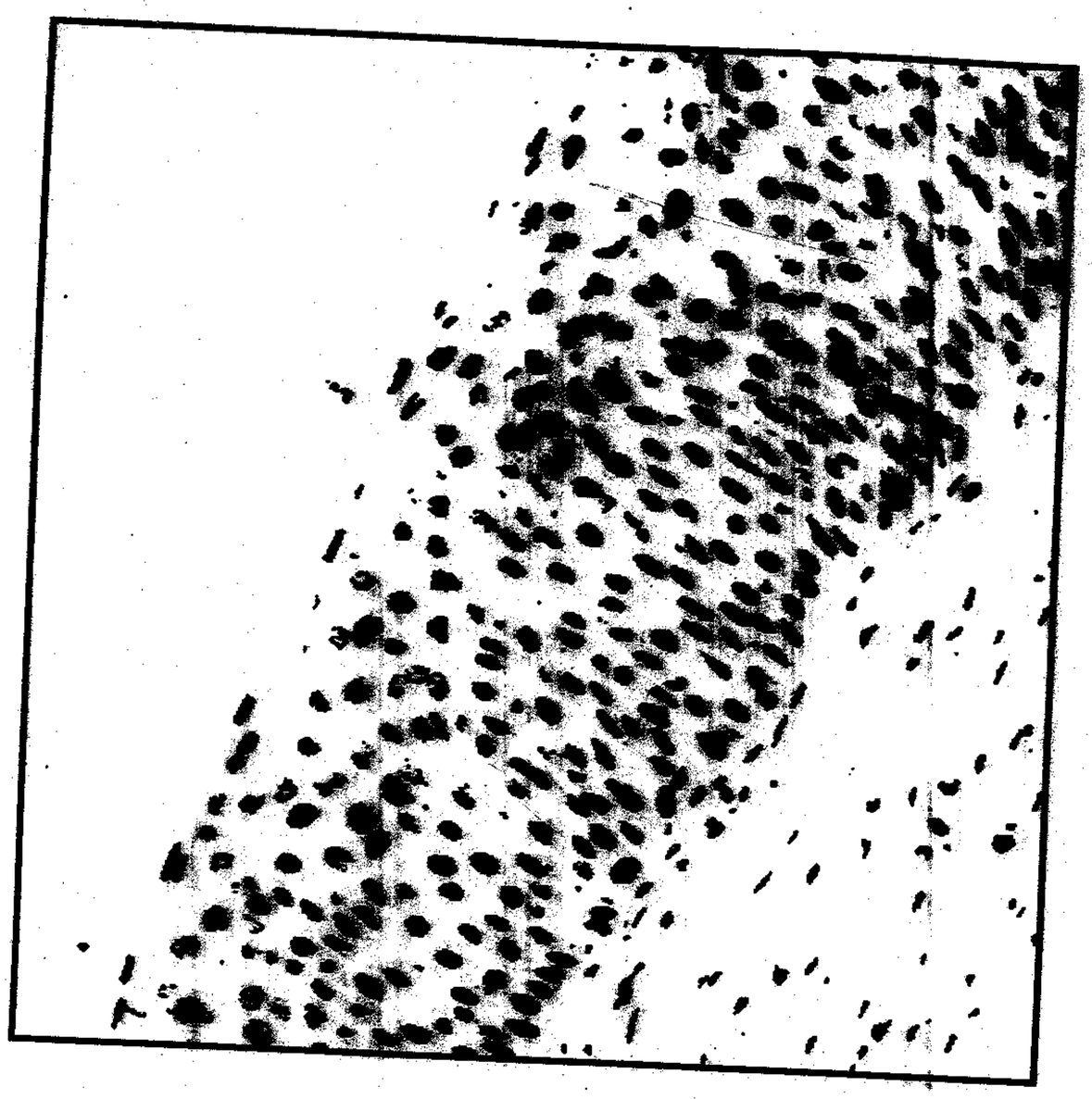


Figure 4.5: Thresholded image with co-occurrence based entropic thresholding algorithm

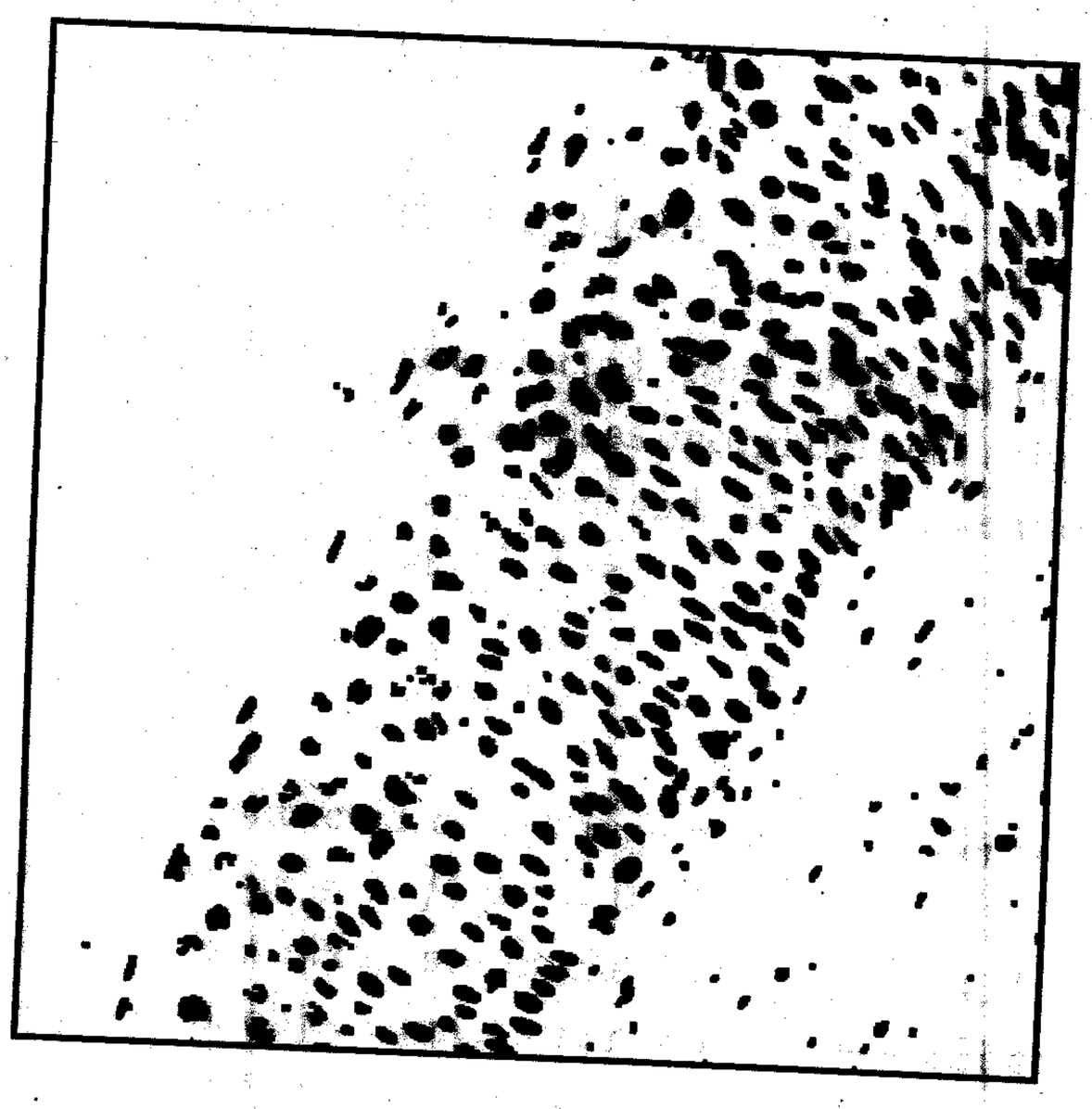


Figure 4.6: Reference image after noise removal

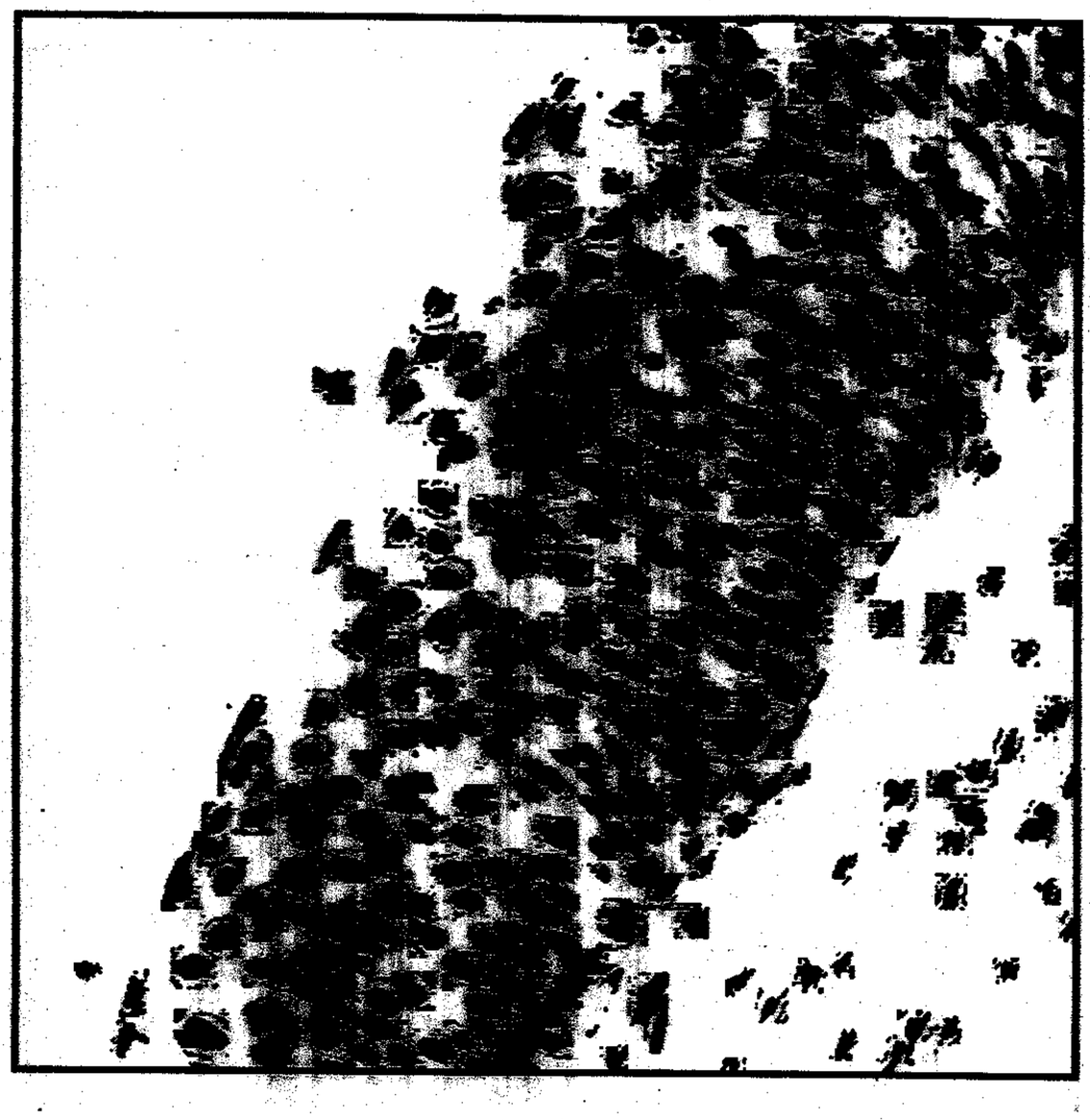


Figure 4.7: Reference image after local thresholding

thresholding algorithm on Fig.4.6 and after conditional dilation operation are shown in figures 4.7 and 4.8, respectively.

4.2.3 Morphological splitting algorithm

Here each individual cell nucleus is extracted from an overlapping object items. For this purpose, a new algorithm, named as morphological splitting algorithm, is proposed based on morphological erosion and dilation operations. The basic concepts of the algorithm are furnished below.

Basic Concepts :

Step 1: Initially, take a connected object region and consider it as the current object. $n = 1$, l=0, $cur = 1$, $f[cur] = 0$.

Assume that the object pixels are marked as 0. (Here n, l, cur and $f[cur]$ indicate the number of object items, the number (level) of iteration (erosion), the object item currently being processed and the level when the object item cur has formed (i.e., got separated) respectively.)

Figure 4.8: Reference image after conditional dilation operation

Step 2: $l = l + 1$.

One erosion operation is applied on the current object item where the deleted pixels are marked as l.

Step 3: Check the number of connected items. In case it is more than one, we keep one connected item as the current object and for each other item as a new object by making $n = n + 1$ and $f[n] = l$. Also make $f[cur] = l$

Step 4: Check whether all the pixels in the current object got deleted or not. If it is not deleted, go back to Step 2. Otherwise (i.e, when all the pixels got deleted), take back the structure of the current object to its formation level i.e, at $f[cur]$ and mark the pixels as (M + cur). Make $cur = cur + 1$. (Here M is a big integer number such as 500).

If $cur > n$, i.e, all the individual object items have been processed, go to Step 5. Otherwise, make $l = f[cur]$ and go back to Step 2.

Step 5: Find $L = \min_{i=1,2,..,n} f[i]$.

Step 6: Apply one dilation operation on the pixels which are marked as more than M. The pixels with marks L are only added to the different items by marking them with the level of the pixel on which the dilation operation is applied.

In case, the new pixel lying in the boundary of two different items (i.e, which can be added to different items) the pixels marked as background.

Step 7: If there are still more pixels left with level L we go back to Step 6. Otherwise, make $L = L$ - 1.

If $L > 0$, we go back to Step 6.

Otherwise stop.

The above algorithm is illustrated on an image diagram shown in Fig.4.9 (a), where all object pixels are marked as 0 and background as blank. After applying two erosion, the object is split into two parts Fig.4.9(b).

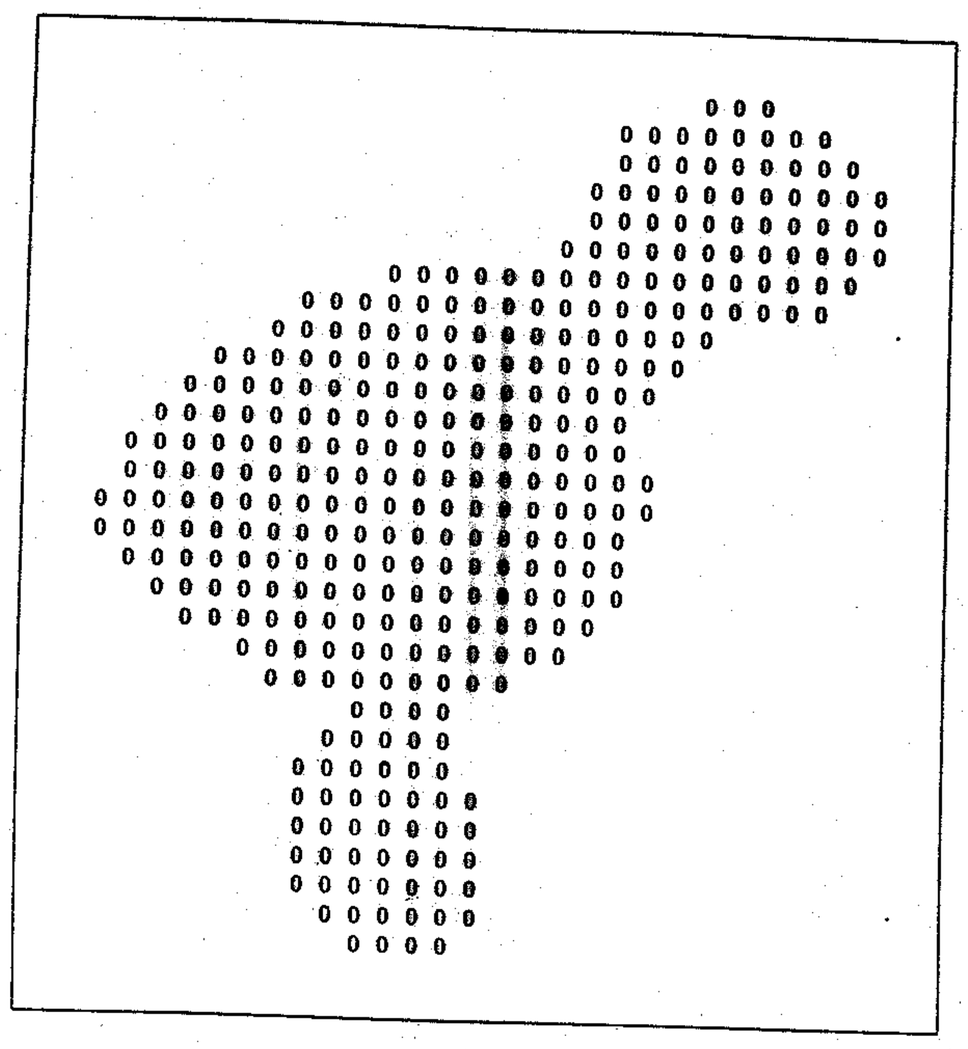
The upper portion is considered now as the current object and the item that lies in the lower side is marked as the second item. After one more erosion the current item again found to be decomposed into two parts [Fig.4.9(c)] by making the top item as the current item and the other one as the third item. One more erosion makes the current object without any object pixels (i.e, 0) and so the pixels which were present in the previous iteration are marked as A [Fig.4.9(d)].

The current object is now taken to be second item i.e, the lower most split object. All the object pixels are found to get deleted after applying one erosion operation and so all of its object pixels are marked as B [Fig.4.9(e)].

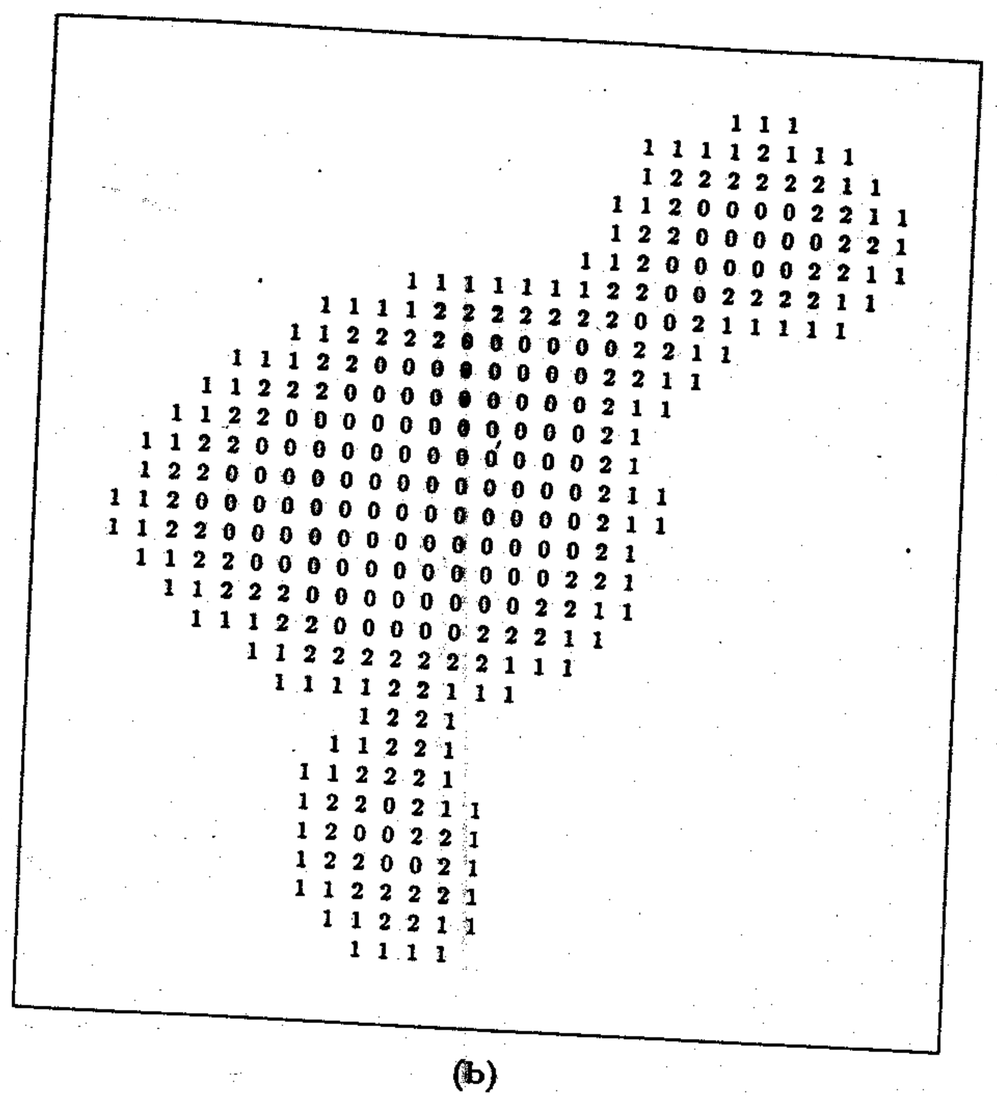
The current object is now taken to the third item i.e, the middle one. After applying three erosions (i.e, up to six operations) all of its object items, are found to get deleted [Fig.4.9(f)]. All the pixels with 4, 5 and 6 are now marked as in [Fig.4.9(g)].

Applying few dilation operations for conditional inclusion of the pixels with value 3 and the pixels with value 2 and finally pixels with value 1, we get three disjoint object items as marked A, B and C in Fig.4.9(h). In each stage of dilation, pixels which are common to more than one object items (i.e., border pixels) are not included.

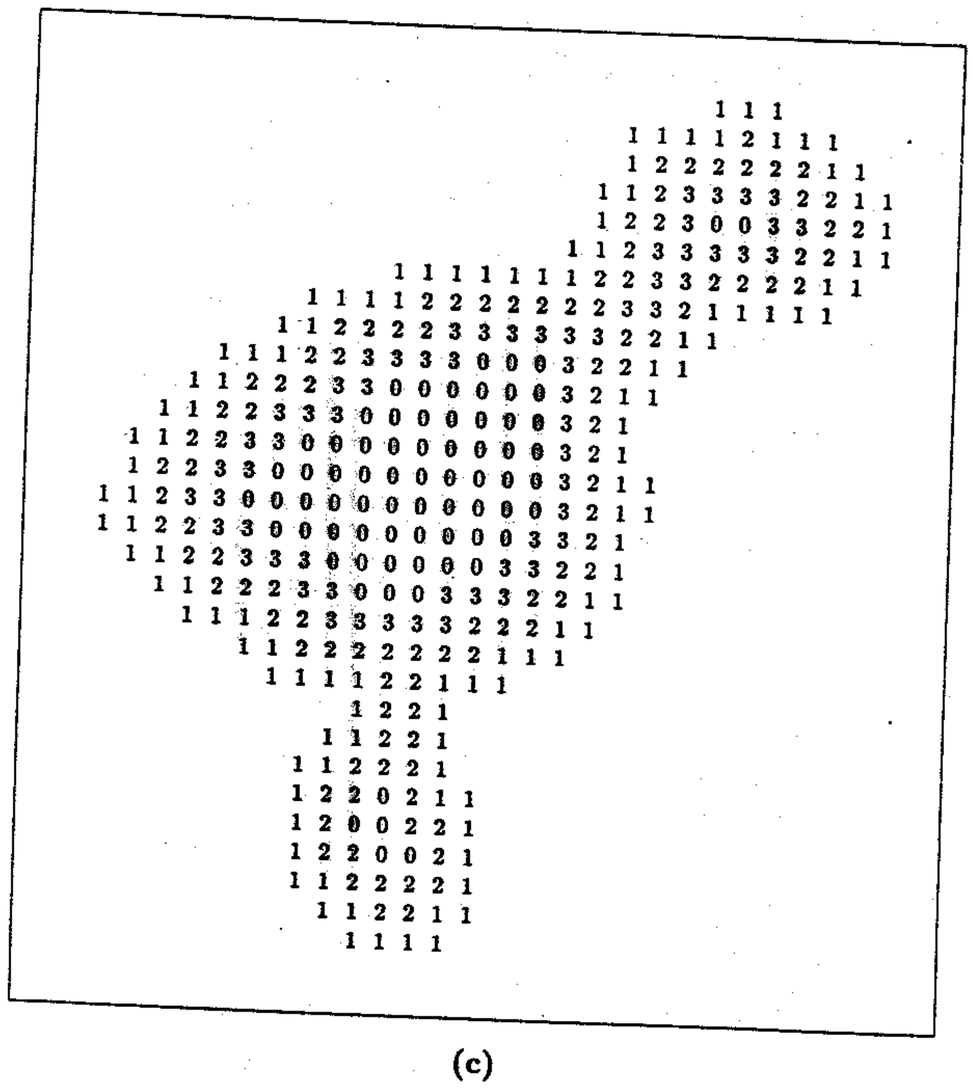
The effect of the morphological splitting algorithm on the image of Fig.4.8 is shown in Fig.4.10.



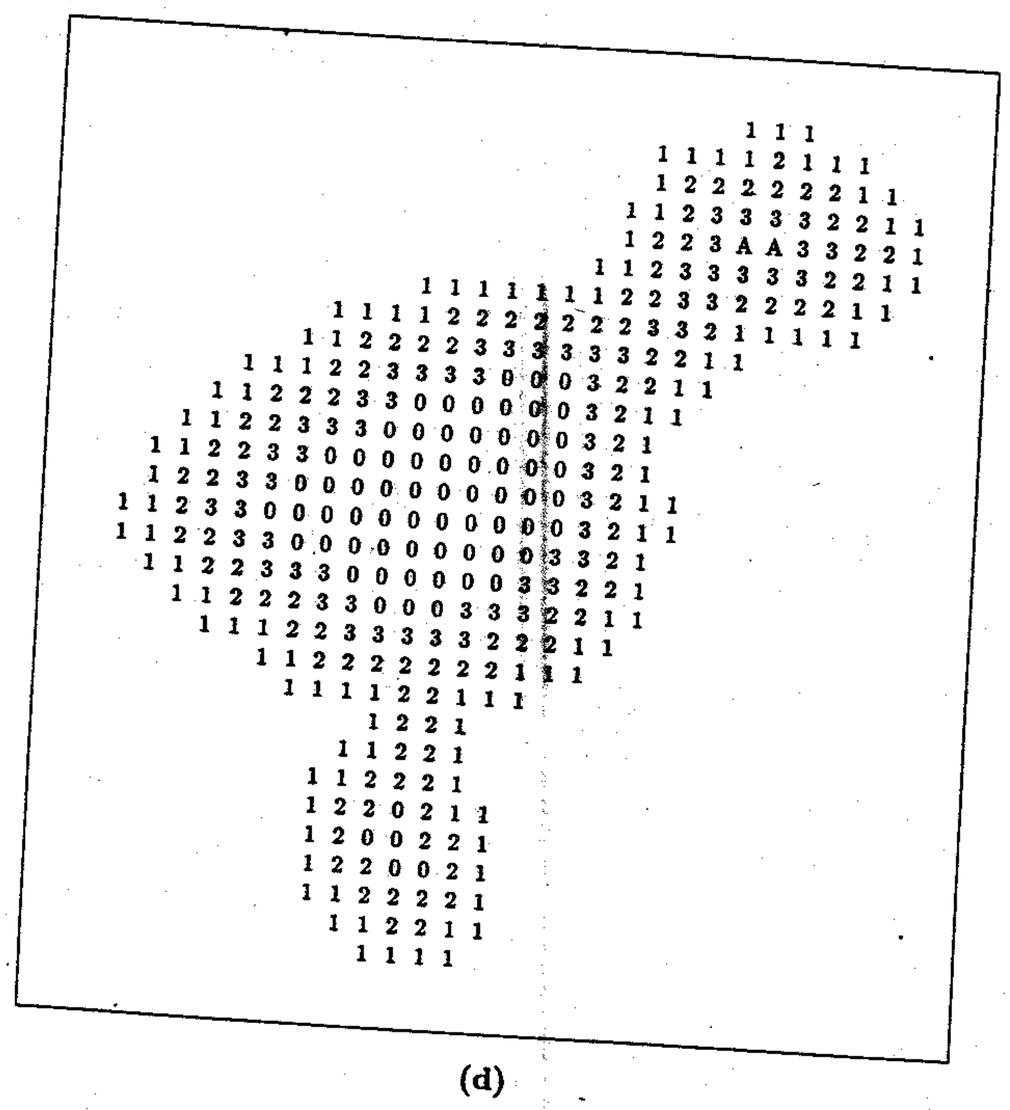
(a)



(b)

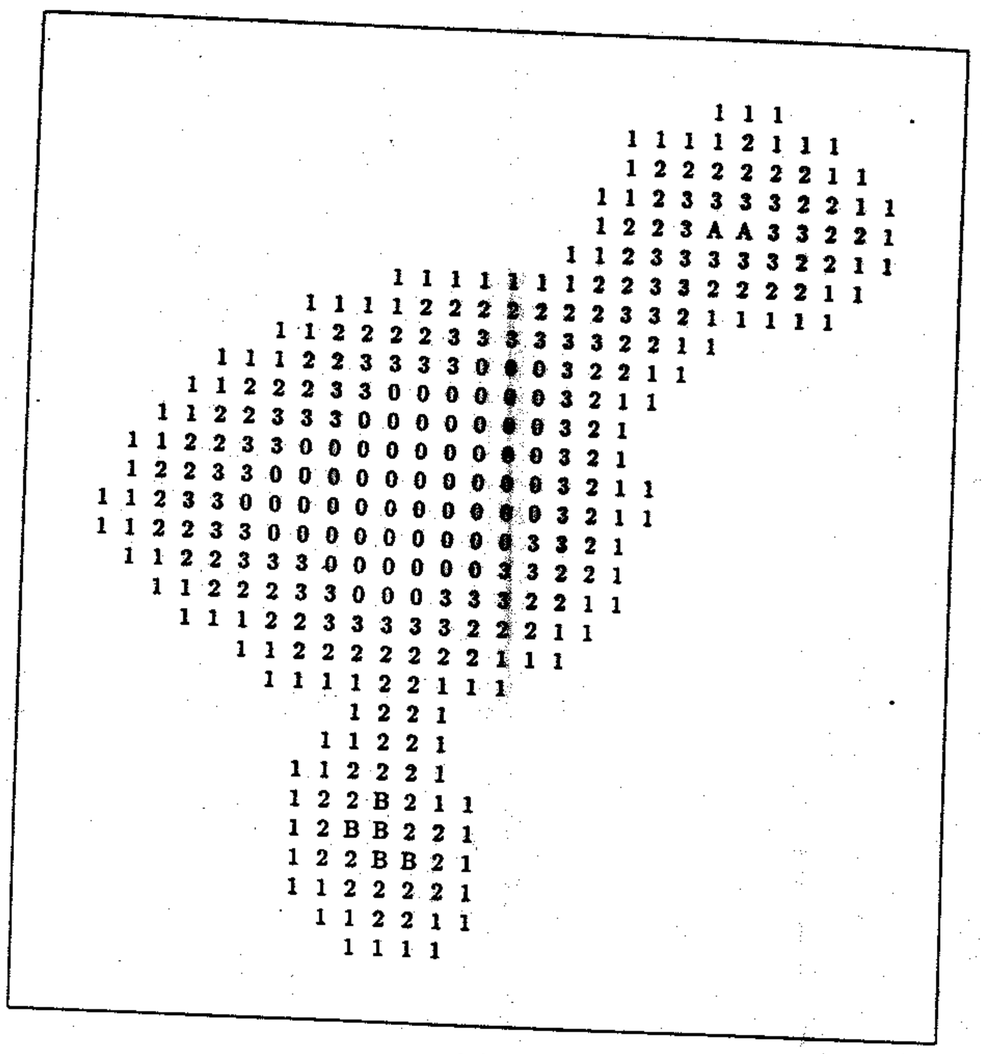


(c)

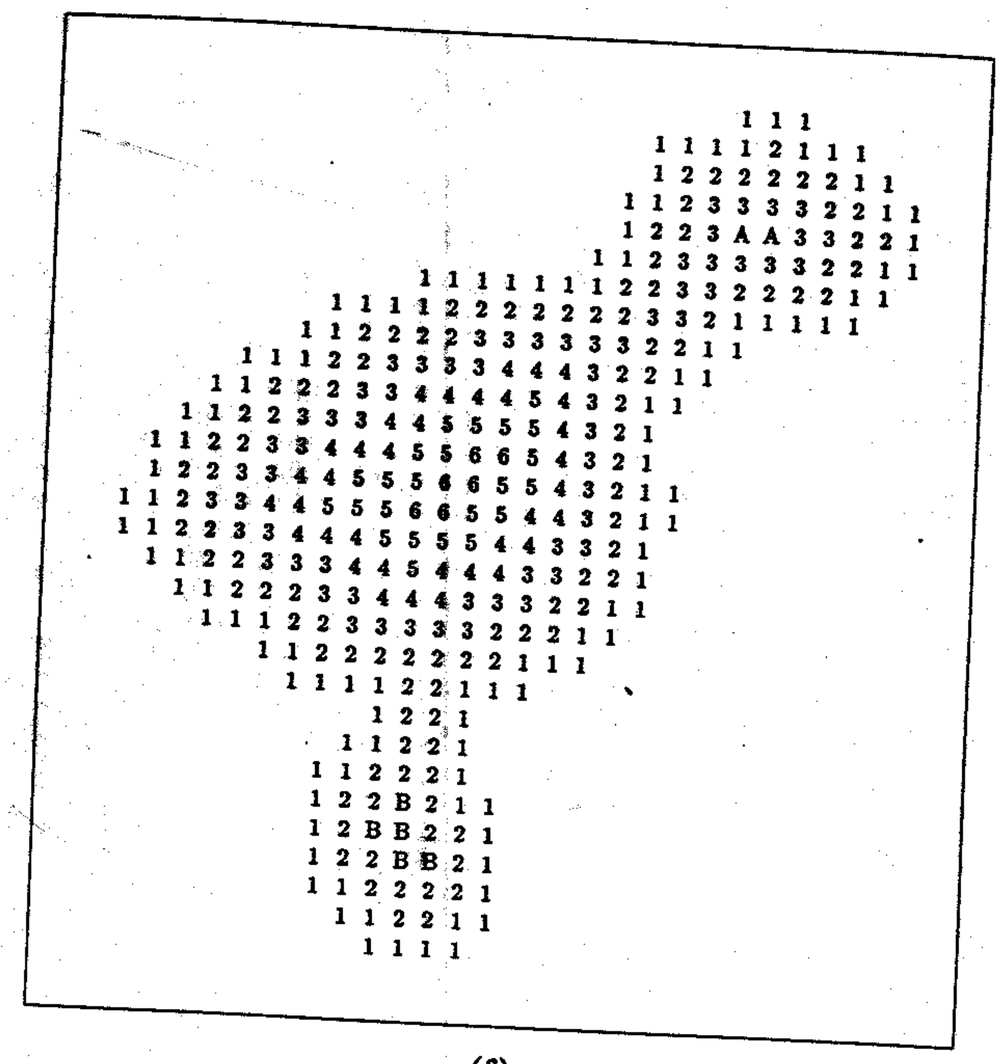


(d)

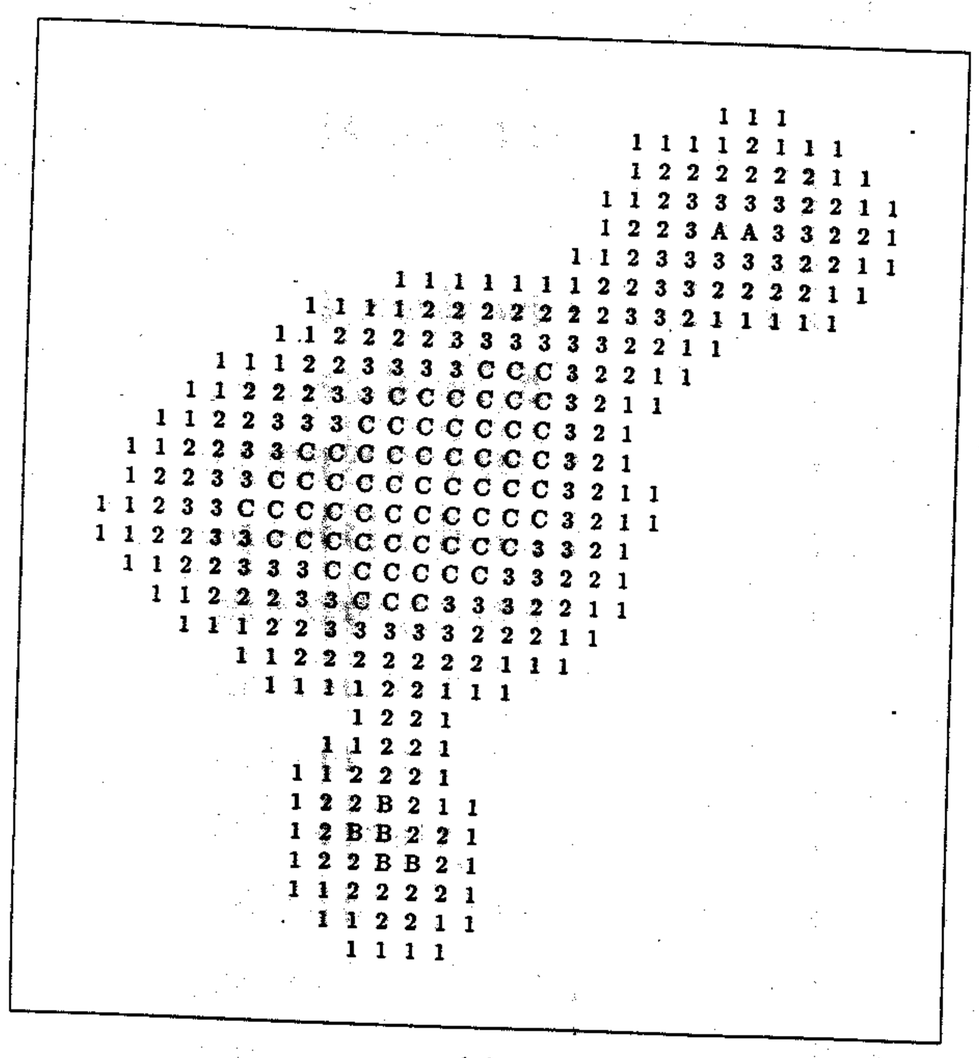
Figure 4.9: Illustrating the morphological splitting algorithm (a) an image diagram, (b) after 2 erosions, (c) after one more erosion on current item, (d) after another erosion on current item, (e) after another erosion on current item, (f) after 3 more erosions on current item, (g) after marking third item and (h) after dilation operations (final output)



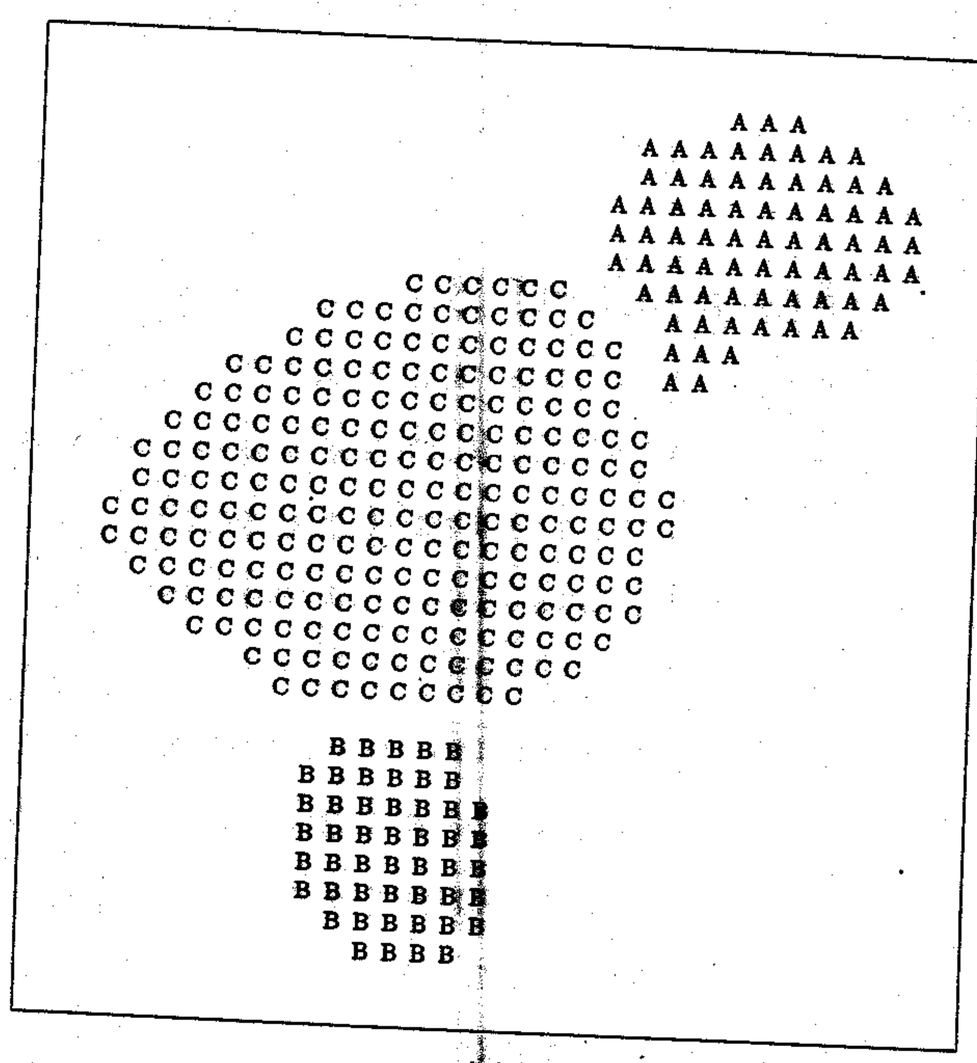
(e)



(f)



(g)



(h)

Figure 4.9: (Continued)

Figure 4.10: Reference image after using the morphological splitting algorithm

4.3 Rotation

This block deals with rotation of the image around its center such that the superficial line is placed horizontally on the top side. The rotation angle is decided based on the more or less homogeneous area with high gray values which usually lies above the superficial line. In the thresholded image, most of the pixels in the homogeneous area are categorized as background. We have utilized here this fact and try to put this area at the top of the image. Accordingly the superficial line is placed horizontally on the top side just below the homogeneous area. On thresholded image, consider a matrix of size 16×16 where each element represents a square area of size 32×32 pixels. The number of object pixels in each square element is counted. All connected blank elements are found and we select the connected blank area which has the maximum size. Then all the elements from the selected blank area which are adjacent to the elements having object pixels are marked. A regression line is fitted based on the marked elements. The slope (θ) of the line is calculated as,

$$tan(\theta) = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i {x_i}^2 - (\sum_i x_i)^2} \tag{4.1}$$

(where, (x_i, y_i) are the ith coordinate of selected blocks, and n is the number of elements marked.)

We will mainly find four different situations as shown in figures 4.11 (a) - (d) and correspondingly, the value of β (rotated angle) is decided as,

Case 1 [Fig.4.11(a)]: When $\theta > 0$ and homogeneous area lies in the left and bottom corner, β = -90 - θ.

Case 2 [Fig.4.11(b)]: When $\theta > 0$ and homogeneous area lies in the top and right corner, β = 90 - θ.

Case 3 [Fig.4.11(c)]: When $\theta < 0$ and homogeneous area lies in the left and top corner, β = -90 - θ.

Case 4 [Fig.4.11(d)]: When $\theta < 0$ and homogeneous area lies in the bottom and right corner, β = 90 - θ.

For the thresholded image of Fig.4.8, we find the rotated image as in Fig.4.12 where the value of β is found to be -64.92 degree.

4.4 Superficial Membrane Identification

On the top side of the rotated thresholded image, we find the homogeneous area with high gray values. When various texture features are computed over a rectangular window around each pixel, the pixels in the homogeneous area exhibit low prominent and shading values in comparison to the pixels in other areas including those on the superficial membrane. Using this observation, this block identifies the superficial membrane.

We applied here the FCM algorithm as described in section 3.1.1. Various window sizes around each pixel are considered for computing several textural feature values (described in section 3.2.3). Based on the two features shading and prominence with 3 classes, the output of the FCM algorithm corresponding to window sizes 7×7, 11×11 and 15×15 are shown in figures 4.13 (a), (b) and (c), respectively. Among these 3 outputs, Fig.4.13 (c) (with window size 15×15) is seen to be better for distinguishing the superficial membrane from the top homogeneous area.

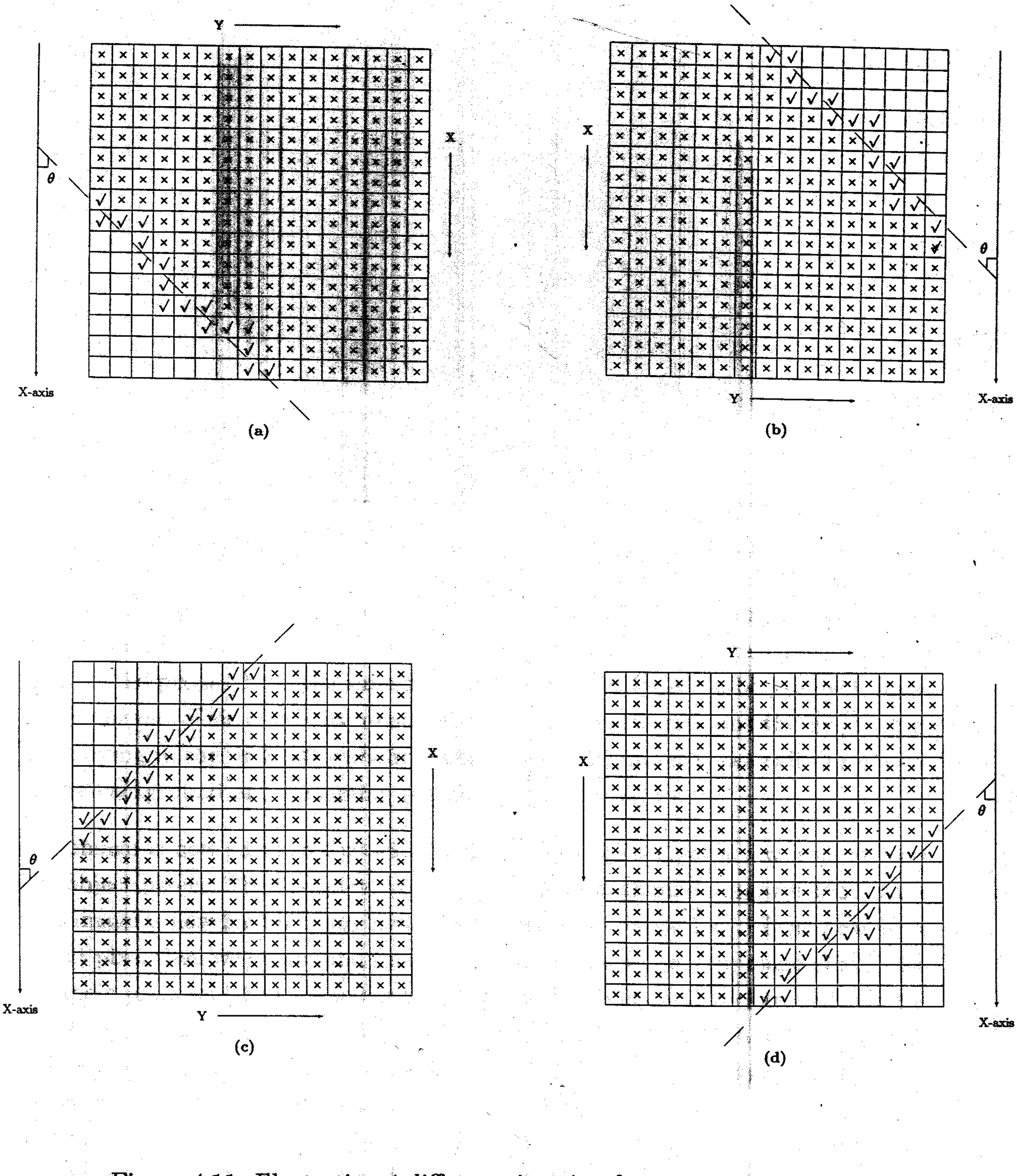


Figure 4.11: Illustrating 4 different situation for rotating an image

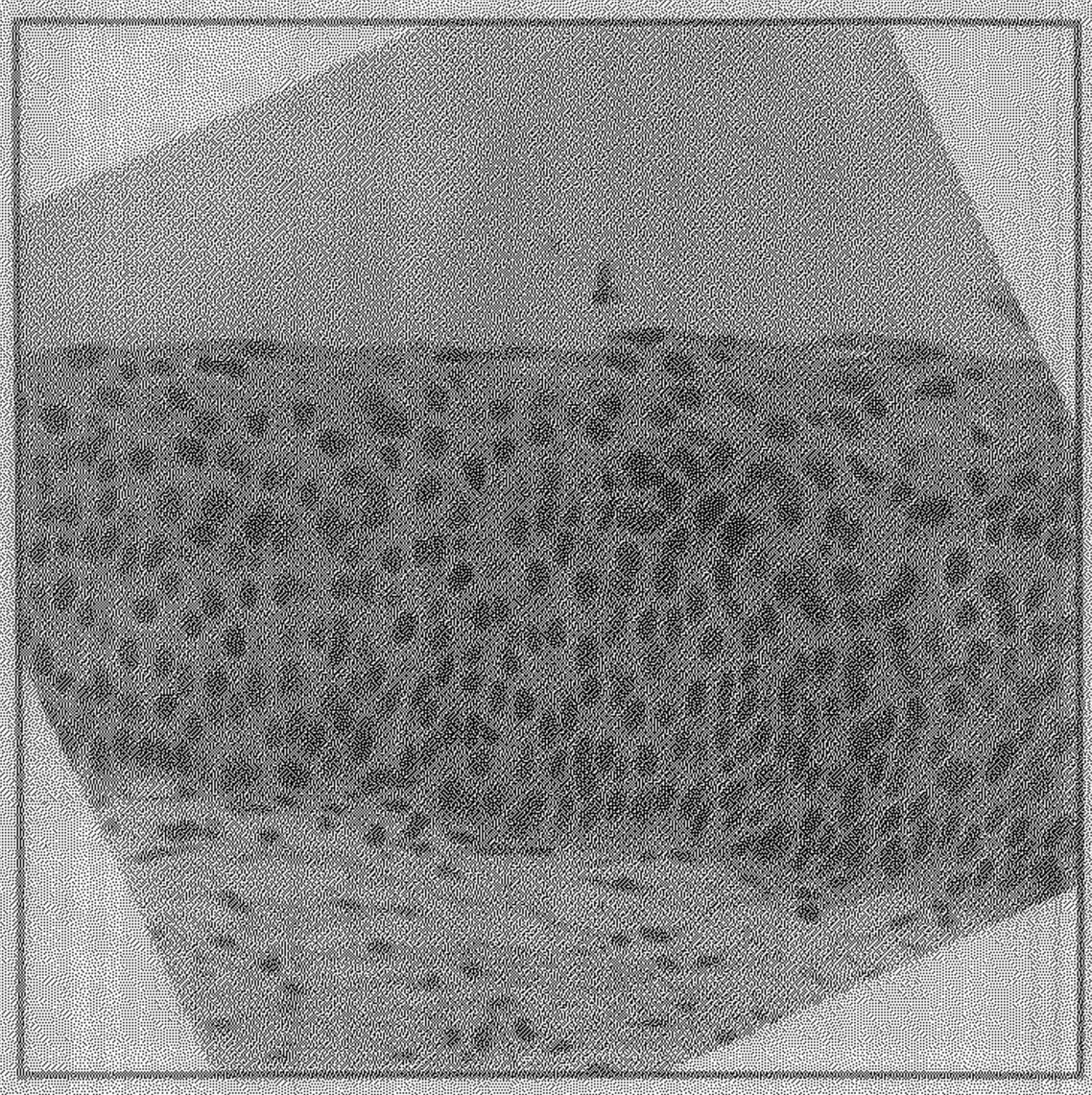
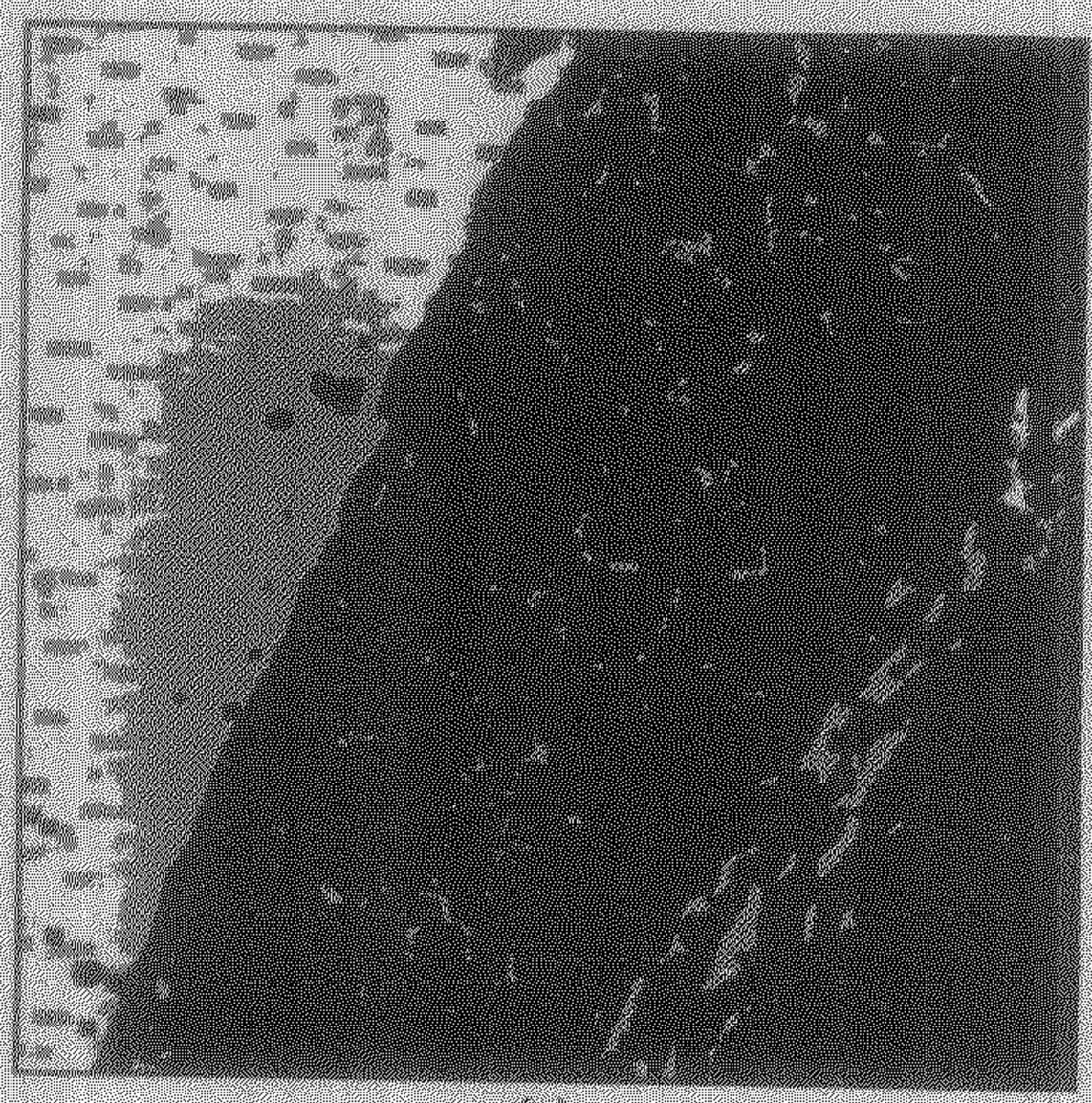


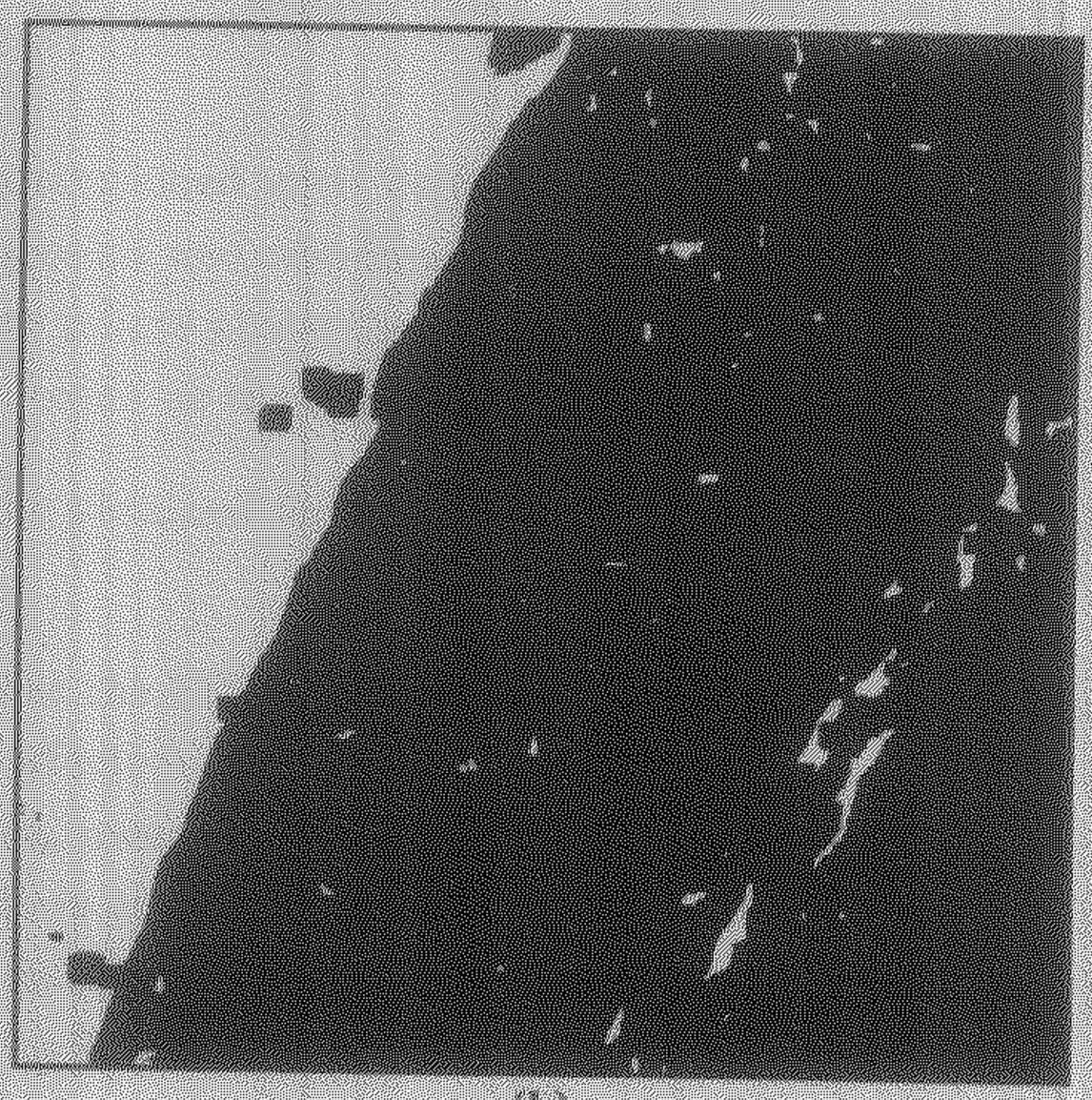
Figure 4.12: Rotated version of the reference image

For identifying the superficial membrane, the output of FCM algorithm with shading and prominence features on 15×15 around each pixel is analyzed. In this image, most of the pixels above the superficial membrane are found to be grouped into a single cluster. This cluster is identified as the one which contains the majority of pixels lying in the first 100 rows from the top. At this moment we call this cluster as the background and remaining two clusters as the object.

The image is now scanned row wise from the top until we get a row with at least 200 object pixels. The first pixel in the row with a image pixel is now found. (Note that image is now having many non-image pixels due to rotation). If the pixel is an object pixel then we go up (and then to right if we reach to a non-image pixel) until we find a pixel that belongs to the boundary of the object and background. On the other hand, if the pixel is a background one, we go down (and then left, if required) until we reach to a pixel lying in the boundary of the object, background and non-image pixel. This is the leftmost point of the superficial membrane. The image is now traversed right sequentially, to find an object point in the each column lying on the boundary of the object and background. During this traversal if the current column point pixel is found to be within 10 rows from the previous column point, we choose this current point as the point of the superficial line. Otherwise, we discard this point and set the current column point same as the previous column point.

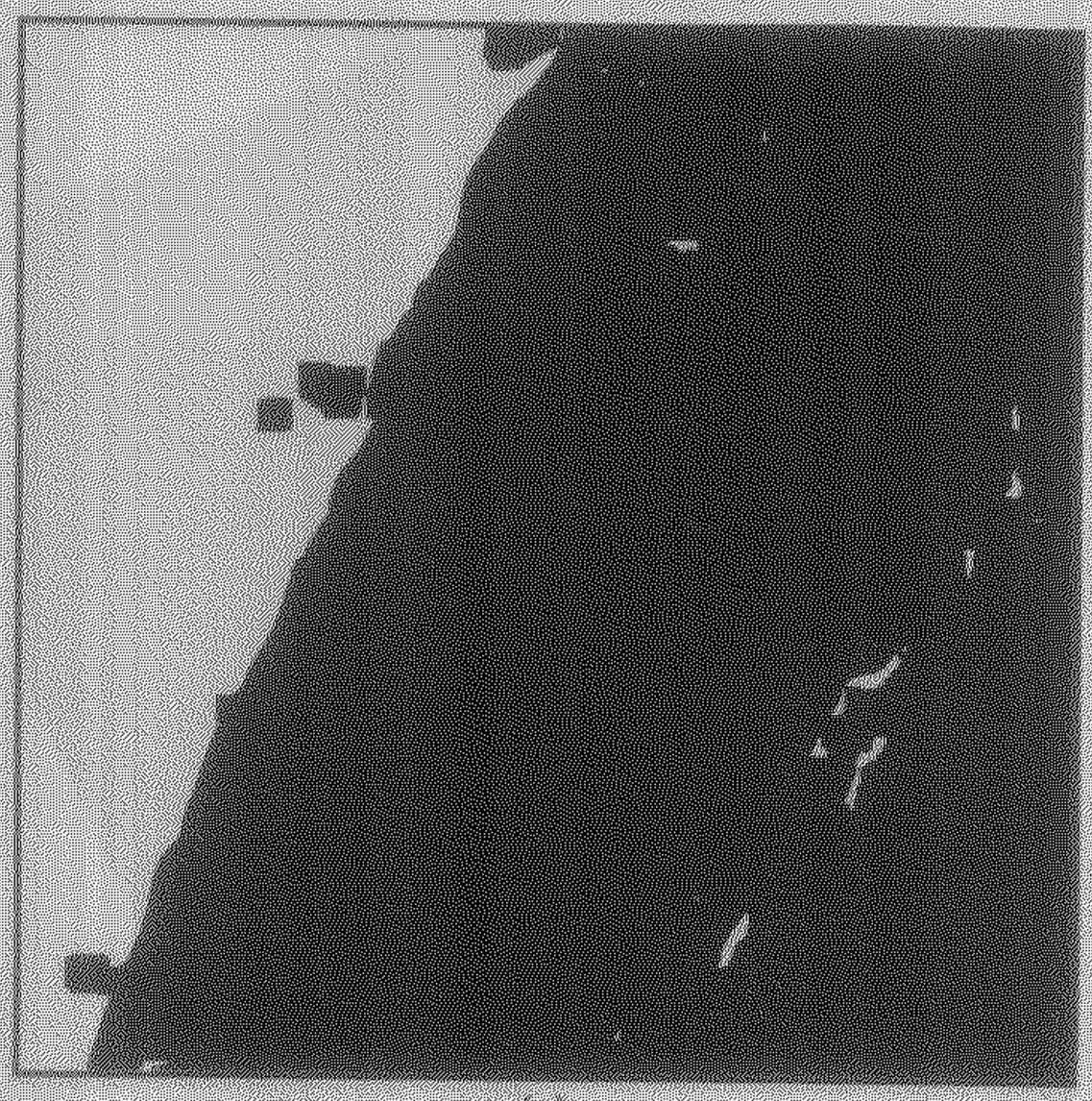


(a)



(b)

Figure 4.13: Reference image after applying the FCM algorithm based on shading and prominence features with 3 classes and window of sizes (a) 7×7, (b) 11×11 and (c) 15×15



(c)

Figure 4.13: (Continued)

This procedure is continued until we find a column point lying in the boundary of the object, background and non-image pixels. This modification is done because there may be some points above the superficial line as object pixels. These column points together provide us the approximated superficial membrane.

The superficial line obtained for the image in Fig.4.12 is shown in Fig 4.14.

Note:

We understand that the FCM algorithm with various other combination of texture features may be effective for slide image analysis. For example, considering the features gray value, mean gray value and standard deviation (s.d.) of gray values for different windows around each pixel, the output of the FCM algorithm with 3 classes for image in Fig.4.2 is shown in figures 4.15 (a), (b) and (c) corresponding to window sizes 7×7, 11×11 and 15×15, respectively.

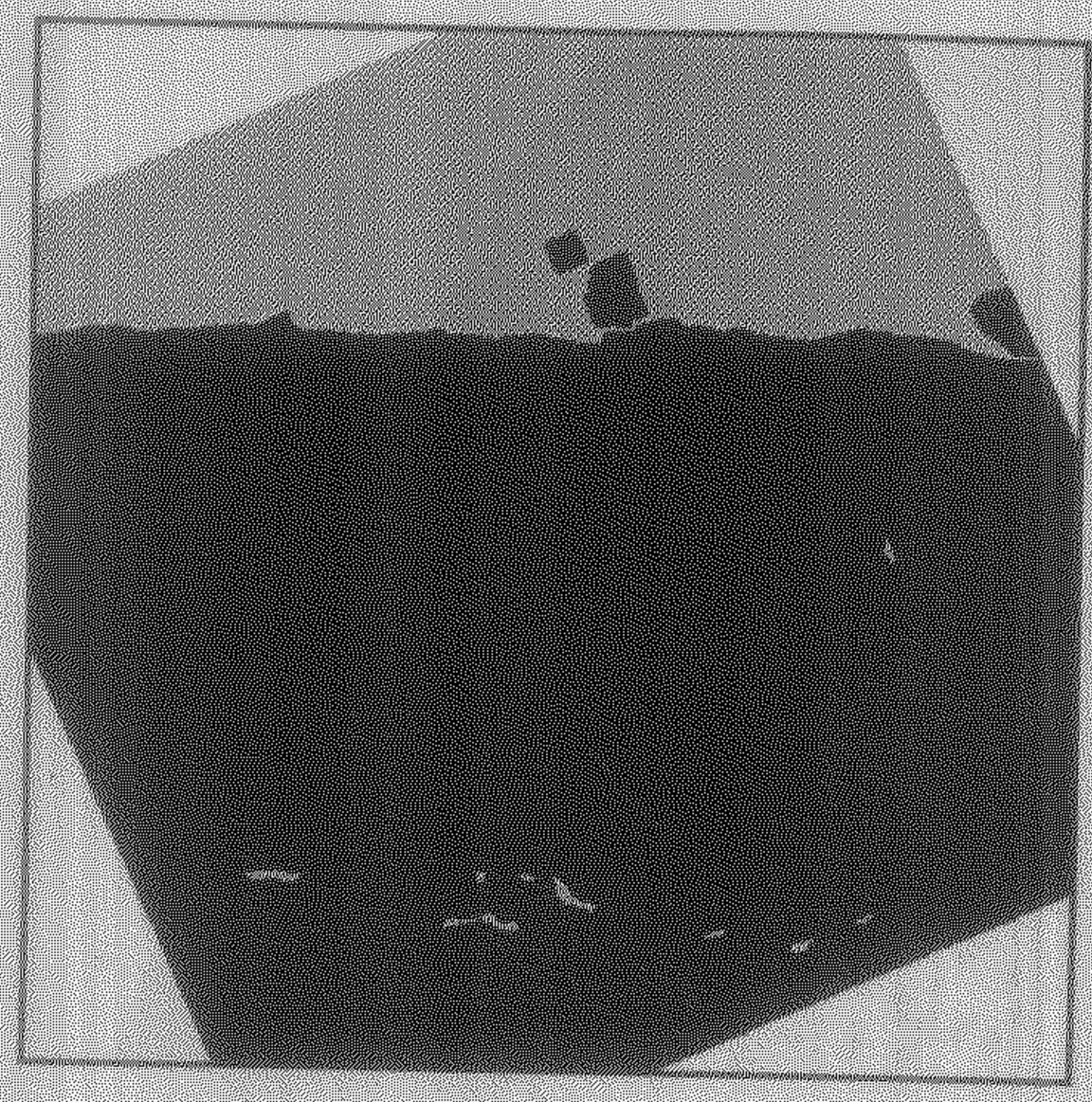
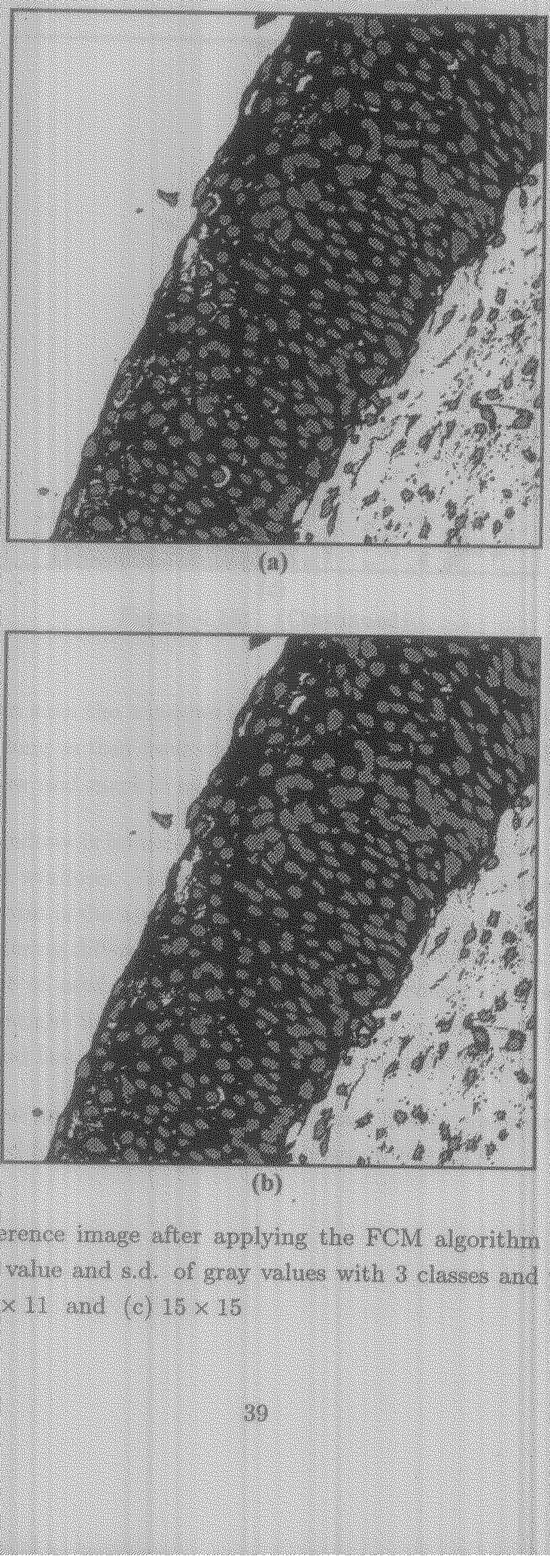


Figure 4.14: Reference image showing the superficial line obtained

4.5 Basal Membrane Identification

It is known that usually the cells just above the basal membrane are vertically elongated and the cells below the basal membrane are horizontally elongated. This block utilizes this observation to identify the basal membrane.

From the improved thresholded image, the horizontally elongated object items are deleted. As a result of this, most of the object items (not nuclei) in the storma regions are likely to get removed. If the height of the smallest rectangular window enclosing an object item is more than its width, the item is considered vertically elongated. Otherwise it is taken as horizontally elongated. The output of the image in Fig.4.12 after removing horizontally elongated items is shown in Fig.4.16. The image is scanned from the bottom column wise, considering overlapping windows of size 32×32. The first column lies between columns 0 and 31 (both included). The first window of this column bounded between rows 480 and 511 (both included). All other windows in this column lie 4 pixel above the previous window, i.e, 28 rows will be common between any two consecutive windows in a column. For example second window in the first column is defined by between rows 476 to 507. Among all the windows in a column, choose the one which has the maximum number of



(a)

(b)

Figure 4.15: Reference image after applying the FCM algorithm based on gray value, mean gray value and s.d. of gray values with 3 classes and window of size (a) 7×7, (b) 11×11 and (c) 15×15

(c)

Figure 4.15: (Continued)

object pixels. Then scan the windows in that column sequentially downwards from the chosen one as long as they have object pixels. Find the lowest object pixel from the scanned window and mark it as a seed point for basal membrane.

Next column of windows is bounded between columns 4 and 35 i.e., 4 pixels right of the first column of windows. Following the above procedure, another seed point is find. If this seed point is the same as previous one, then it is not stored. Proceeding similarly by considering different columns of windows (4 pixels right of the previous one), we get a set of seed points. Then every two consecutive seed points are connected by a straight line. To complete the line, the first seed point extended left horizontally and last seed point is extended right horizontally.

We take the line as an approximation of the basal line or membrane. The basal membrane obtained for image in Fig.4.16 is shown in Fig.4.17.

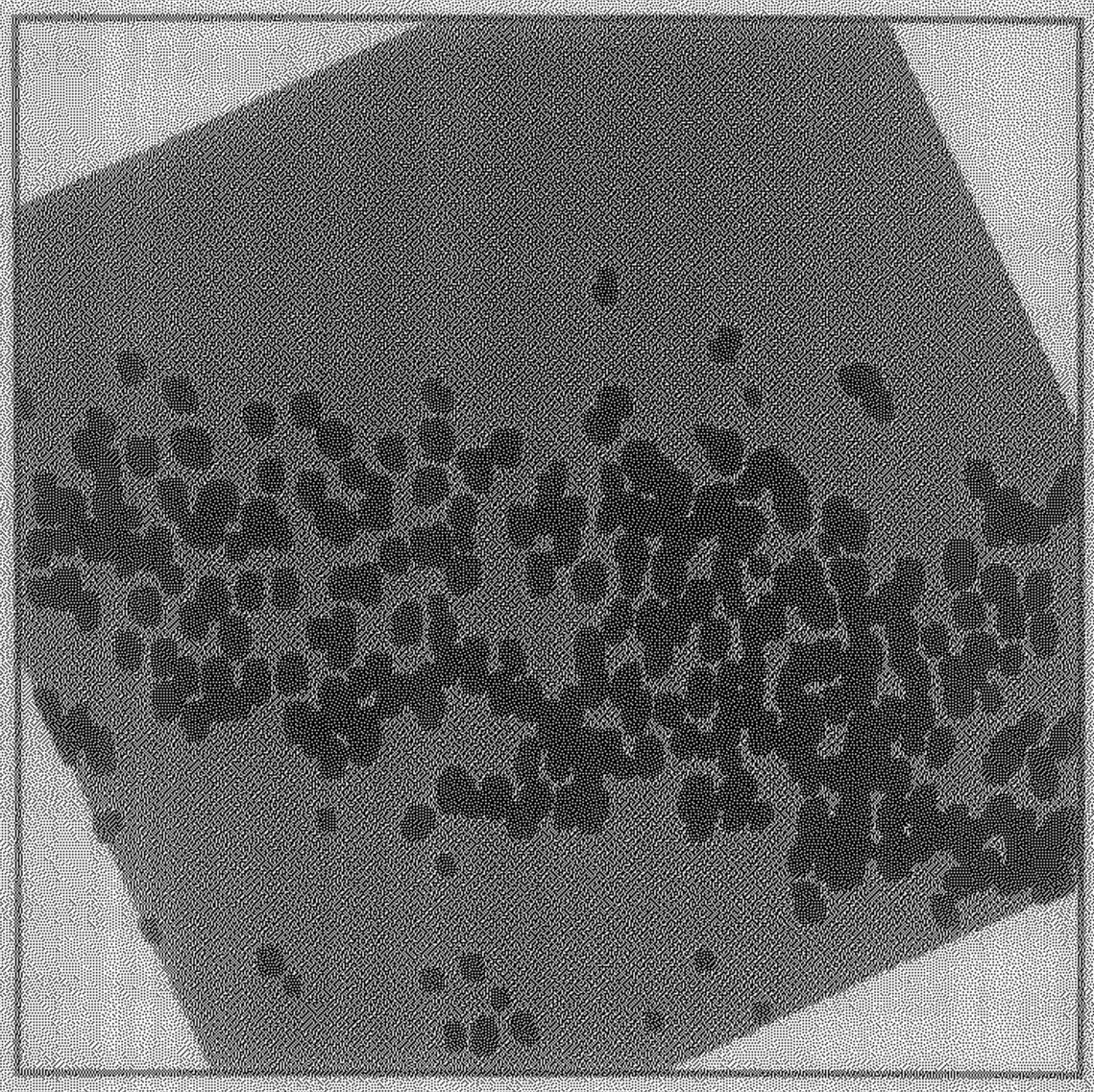


Figure 4.16: Reference image with only the vertically elongated objects

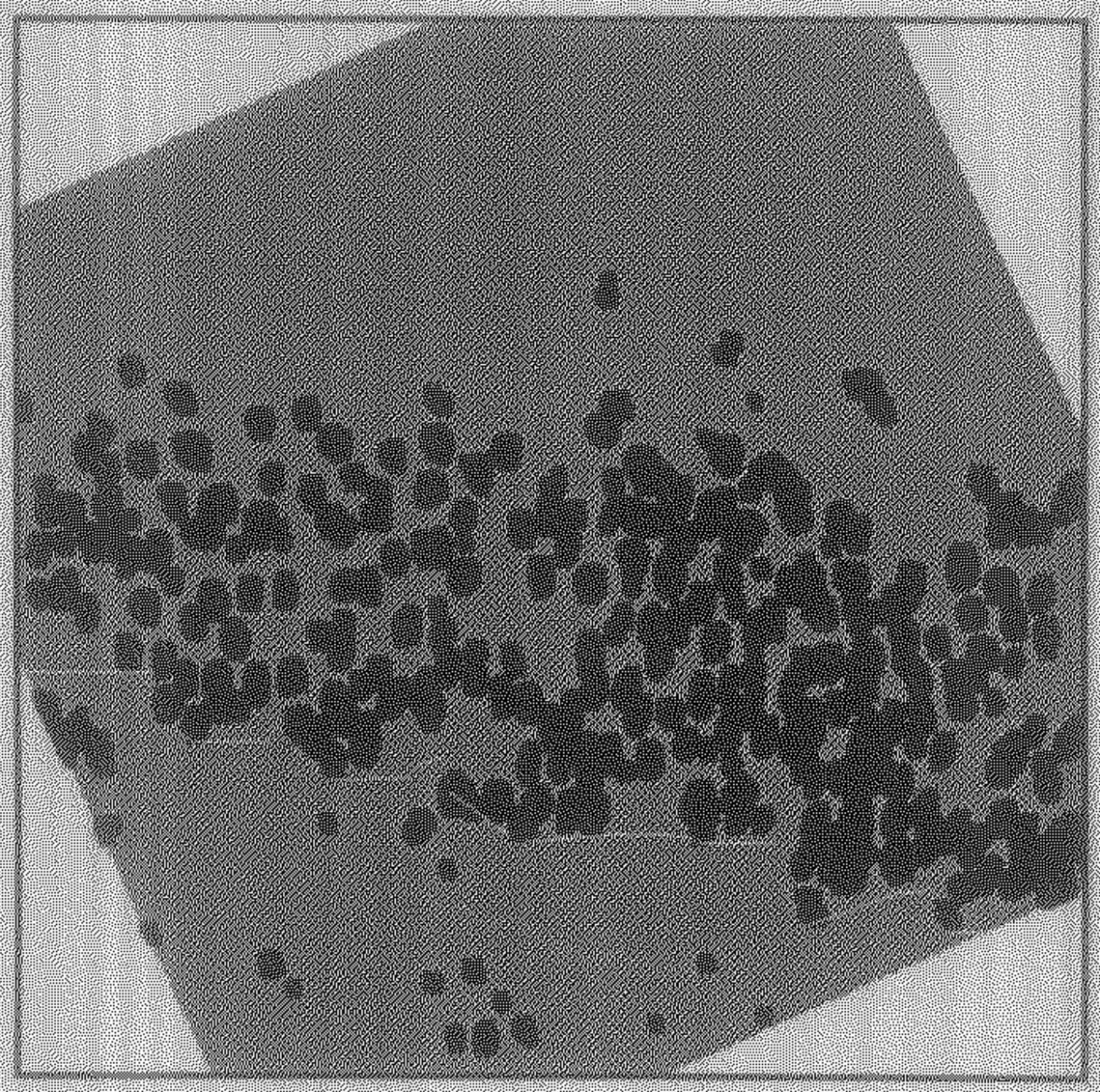


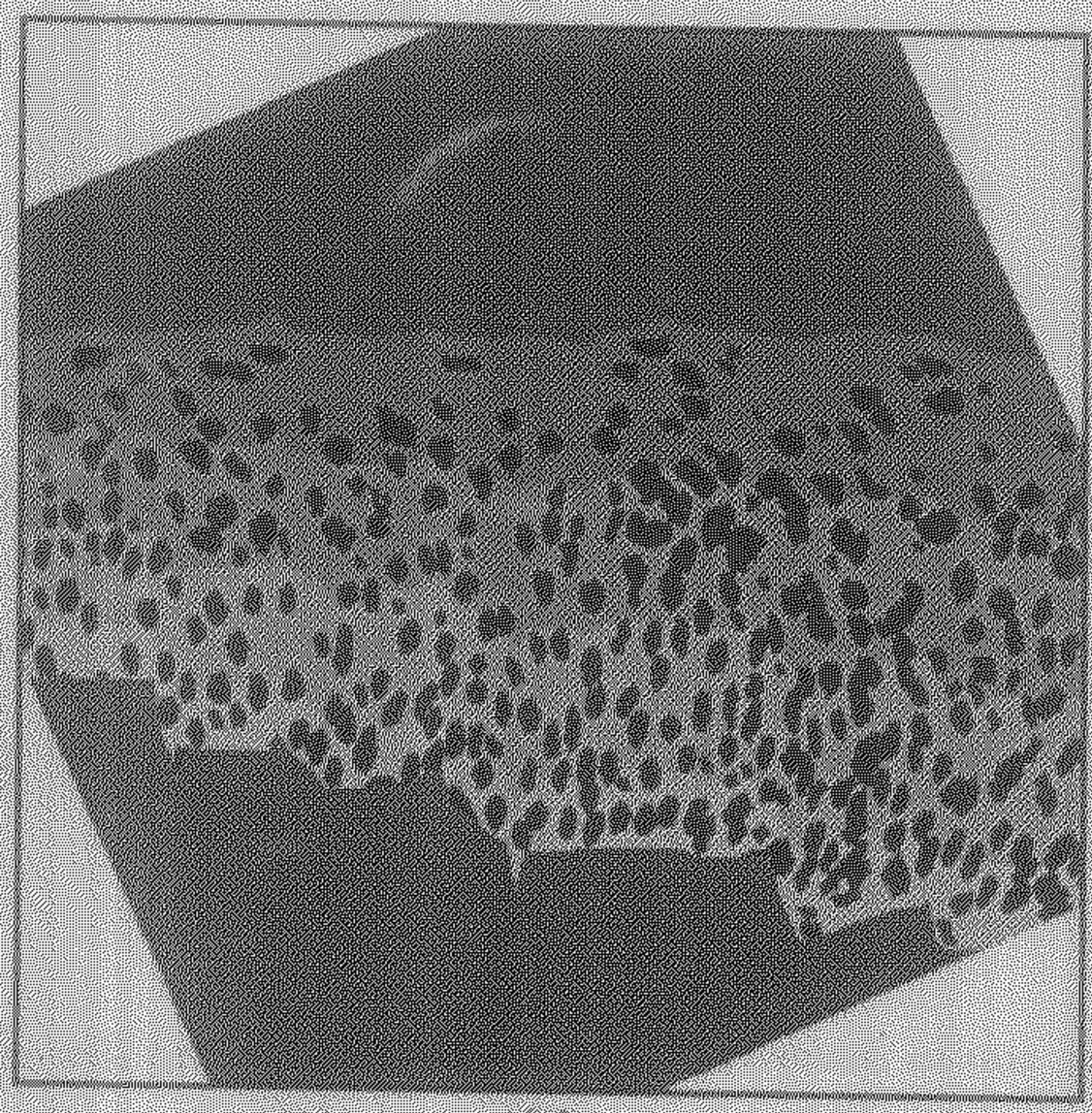
Figure 4.17: Reference image showing the basal line

4.6 Region Identification

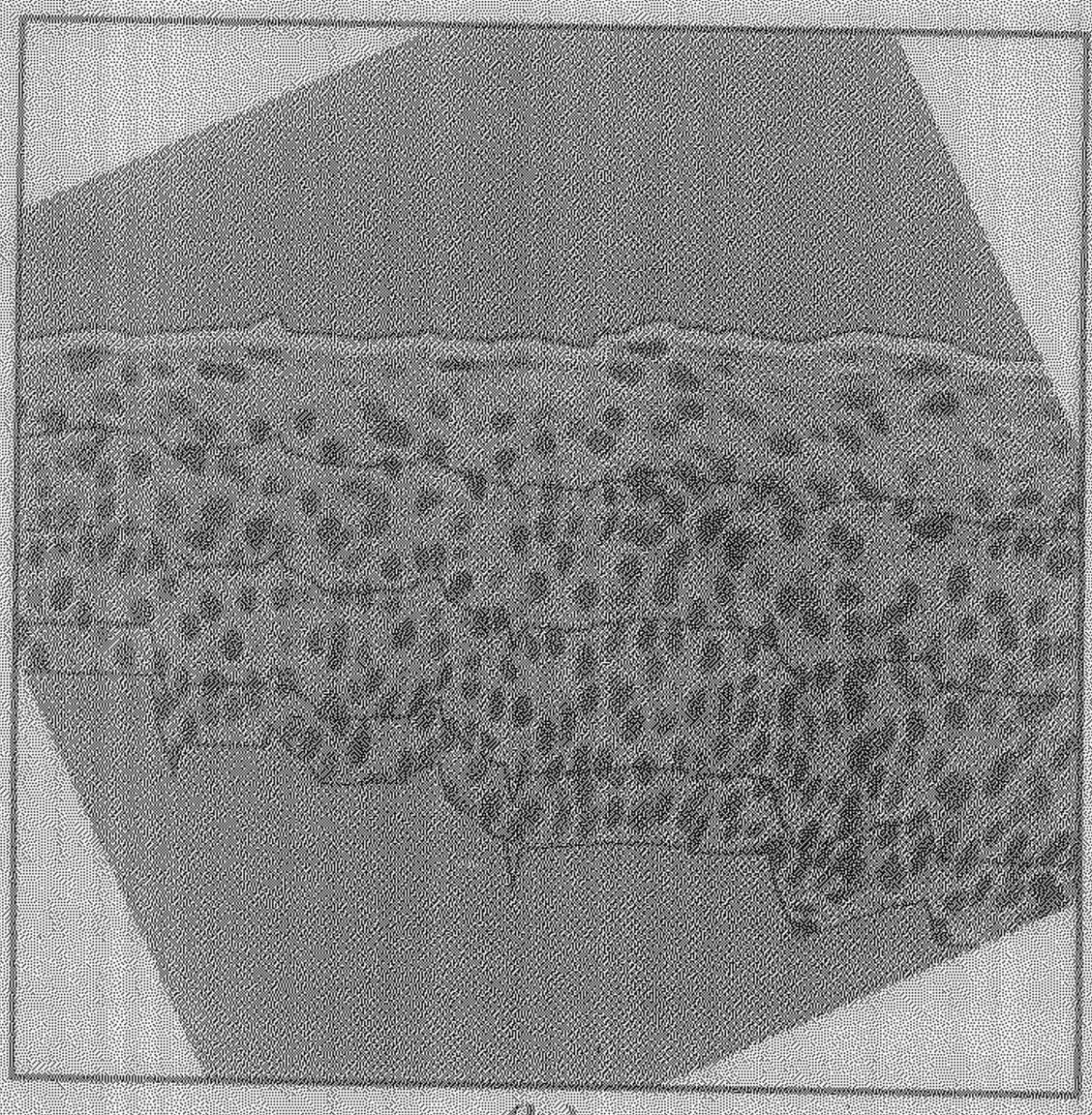
In the previous two blocks, we have marked basal and superficial membranes. The present block decomposes the area between these two membranes into four regions. The area between basal and superficial lines is only analyzed for detecting pre-cancerous cervical changes. This area is composed of four region namely, basal, parabasal, intermediate and superficial regions lying horizontally one below next one in the ratio of 1:2:2:2. The basal and superficial lines are extended horizontally left and right along the non-image area depending upon the start and end column points respectively. For each column, we find 3 points which divides the rows in that column between the basal and superficial lines in the ratio of 1:2:2:2.

As a result we get the basal, parabasal, intermediate and superficial regions. The shaded four regions along with the nuclei only within the regions are shown in Fig.4.18(a). The boundaries of the regions along with original gray values with in the regions are furnished in Fig.4.18(b).

The effectiveness of the proposed strategy is demonstrated on several slide images in the next chapter.



(a)



(b)

Figure 4.18: Showing basal, parabasal, intermediate and superficial regions for the reference image (a) with only the nuclei within the regions, (b) with original gray values within the regions

Chapter 5

Results and Classification

In this chapter we will first demonstrate the effectiveness of our proposed methodologies for the analysis of slide images. The textural feature values obtained corresponding to the extracted basal, parabasal, intermediate and superficial regions are provided. Finally, a strategy for the possible classification is mentioned.

5.1 Results

We have verified the effectiveness of our image analysis approaches (as described in the previous chapters) on many slide images. The results are found to be satisfactory in all the cases. We include here only eight of them (2 from each of the classes normal, CIN-1, CIN-2 and CIN-3). The test images are shown in figures 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7 and 5.8. The first two test images (i.e., Test Image - 1 and Test Image - 2) belong to the class normal; Test Image - 3 and Test Image - 4 are CIN-1; Test Image - 5 and Test Image - 6 belong to CIN-2; and the last two images (i.e., Test Image - 7 and Test Image - 8) lie in the class CIN-3. The basal, parabasal, intermediate and superficial regions obtained are shown in Fig.5.9(a) (along with nuclei) and Fig.5.10(b) (relevant portion of the original image with region boundaries superimposed) corresponding to image in Fig.5.1. Similarly, the regions corresponding to images in figures 5.2, 5.3, 5.4, 5.5, 5.6, 5.7 and 5.8 are shown in figures 5.10, 5.11, 5.12, 5.13, 5.14, 5.15 and 5.16.

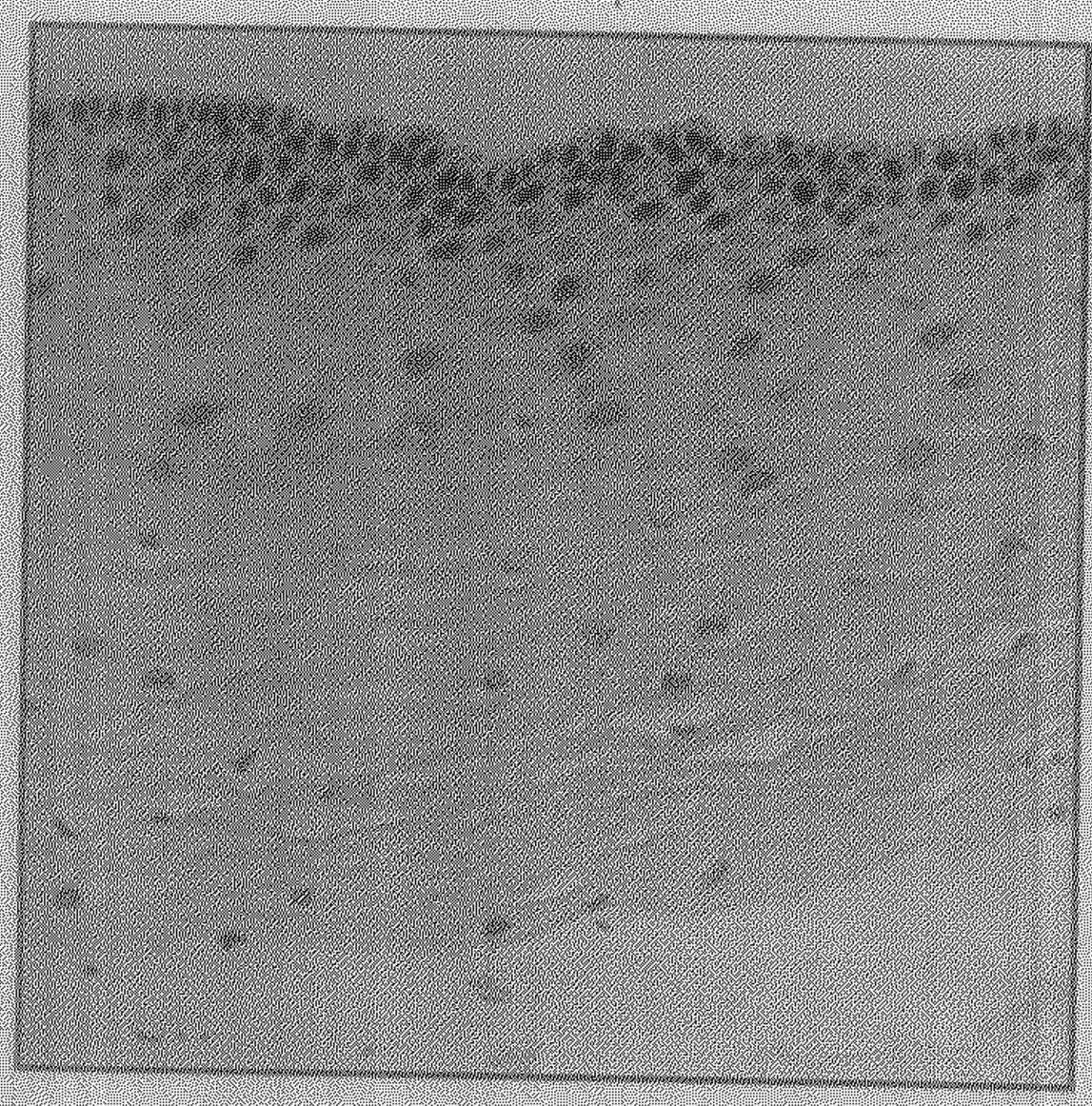


Figure 5.1: Test Image -1

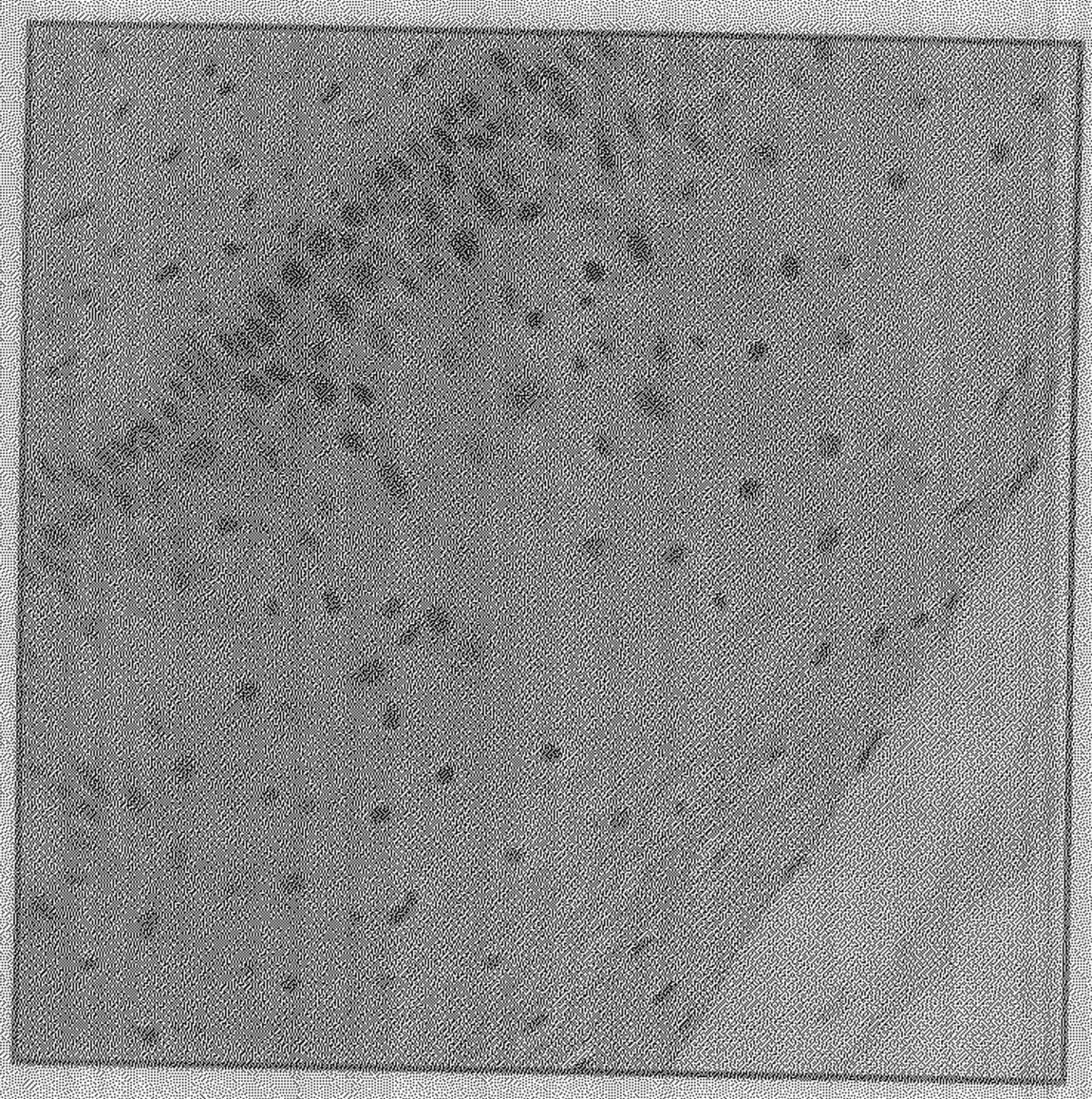


Figure 5.2: Test Image -2

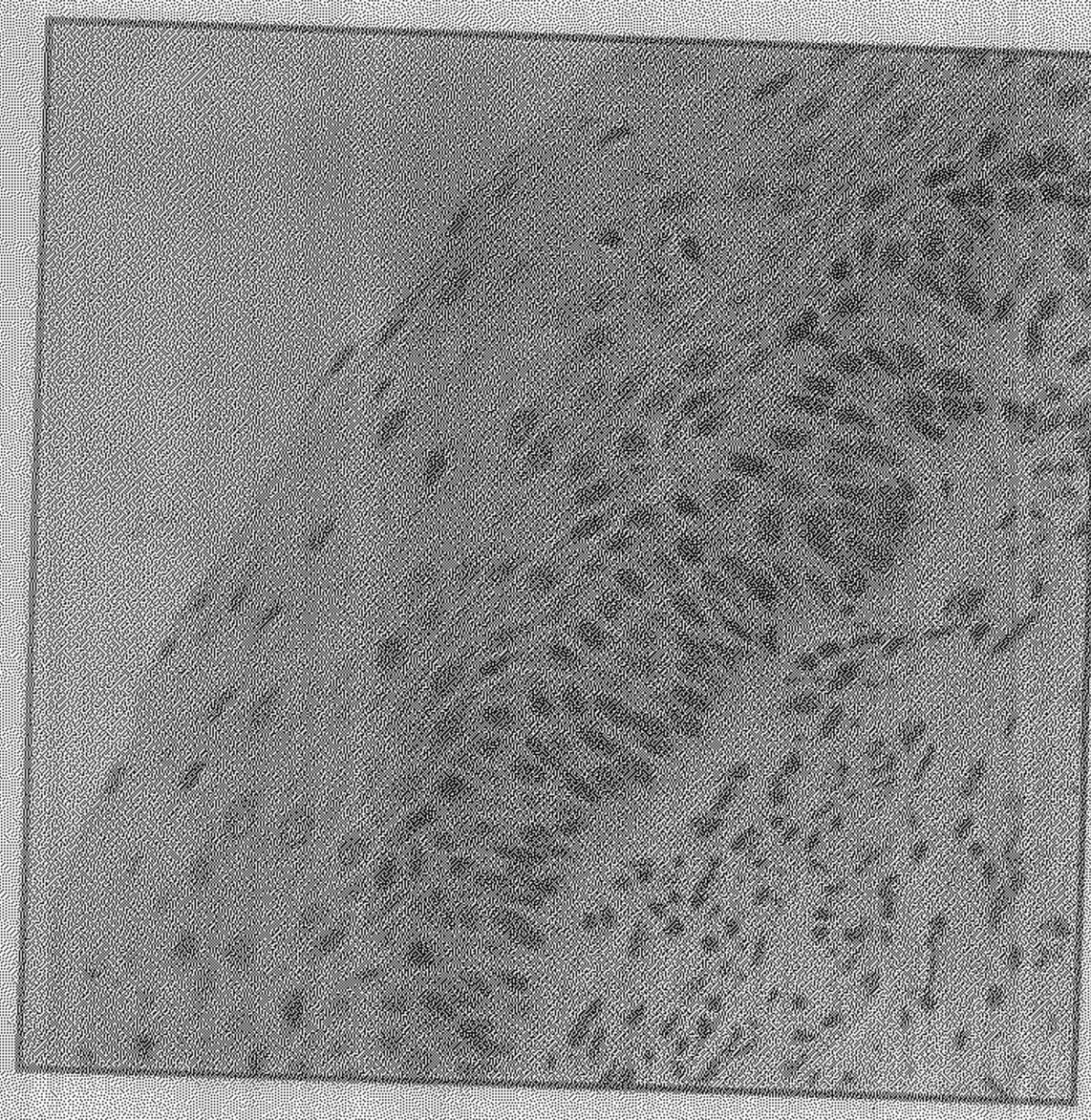


Figure 5.3: Test Image -3

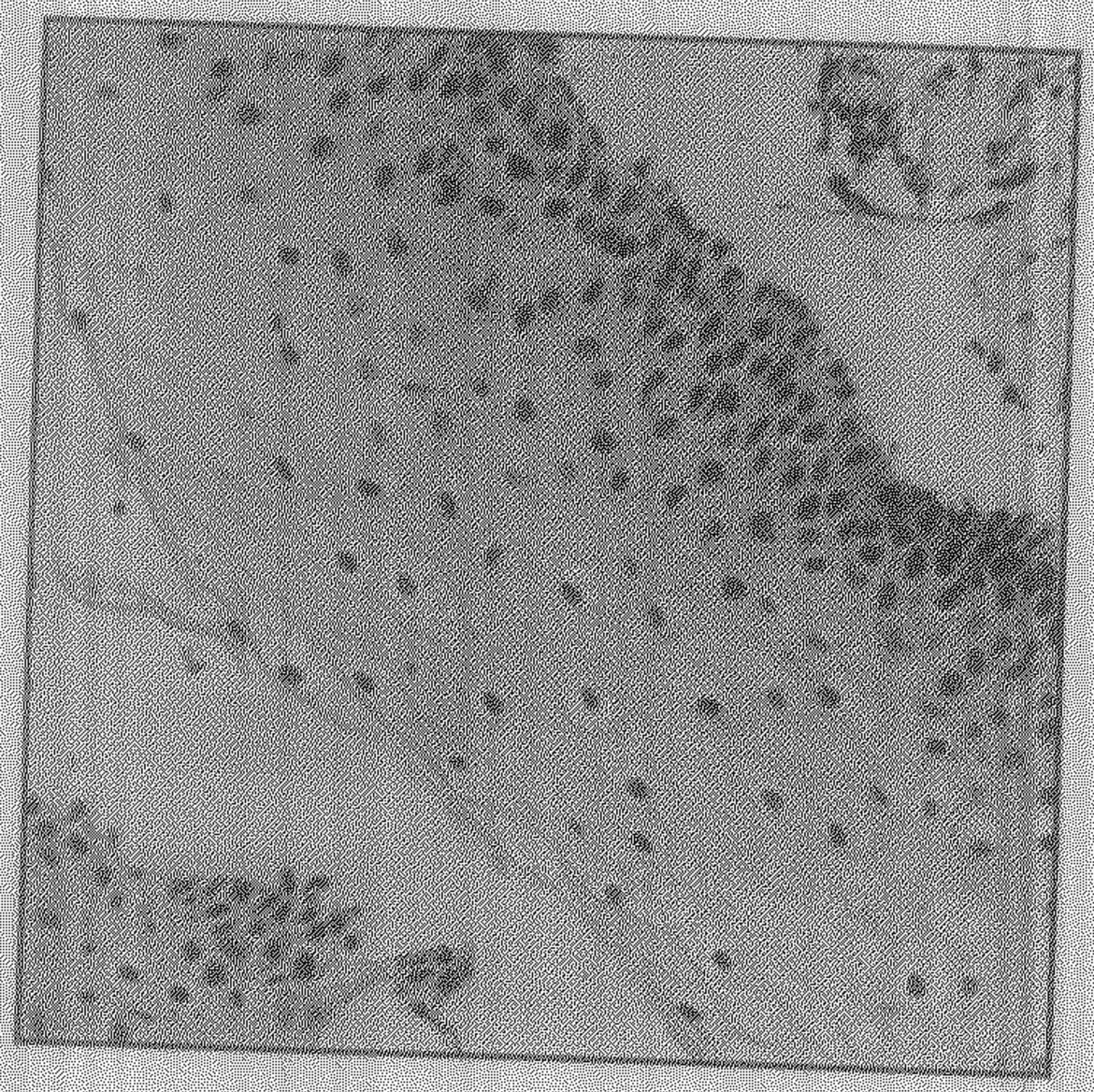


Figure 5.4: Test Image -4

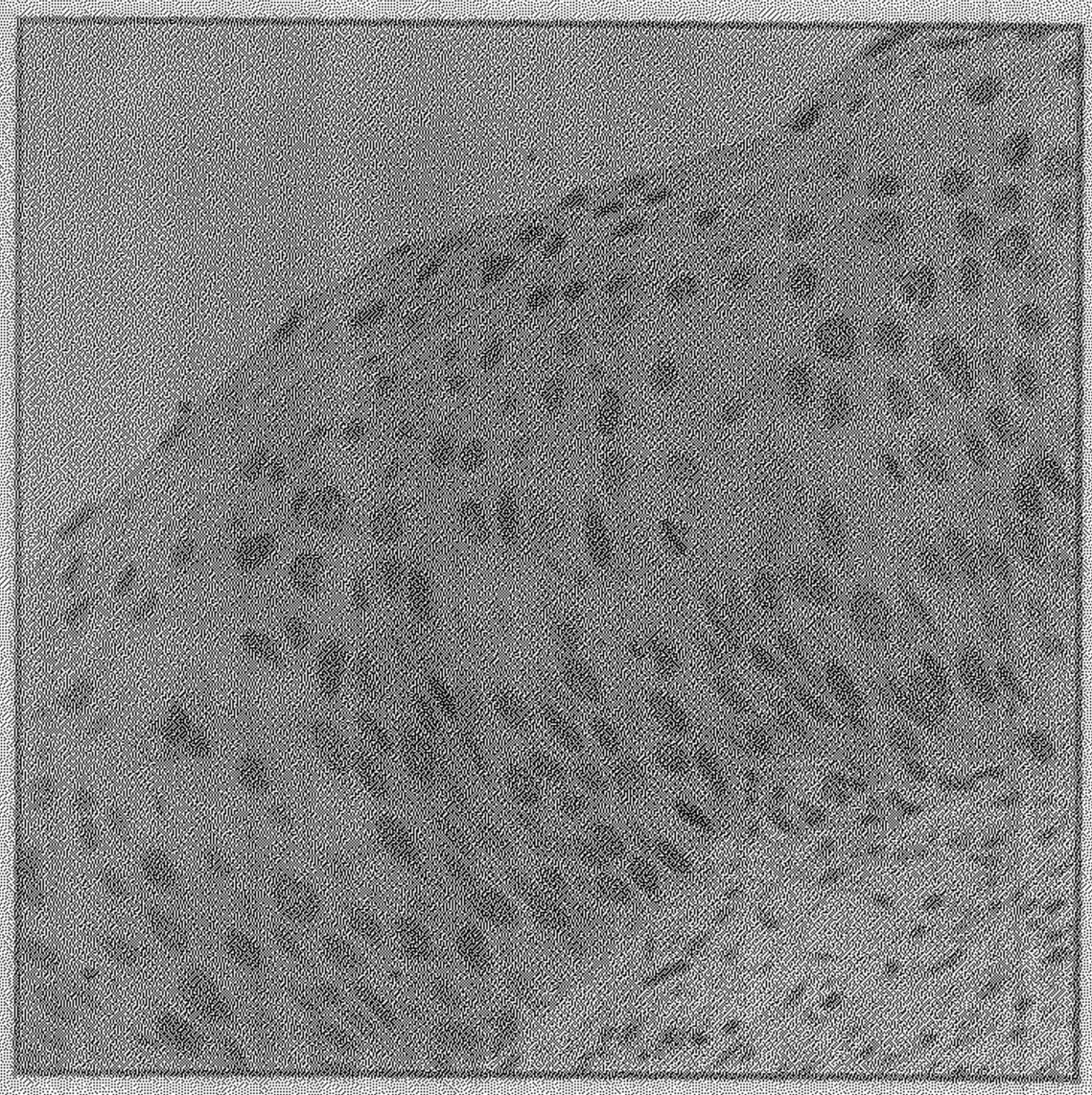


Figure 5.5: Test Image -5

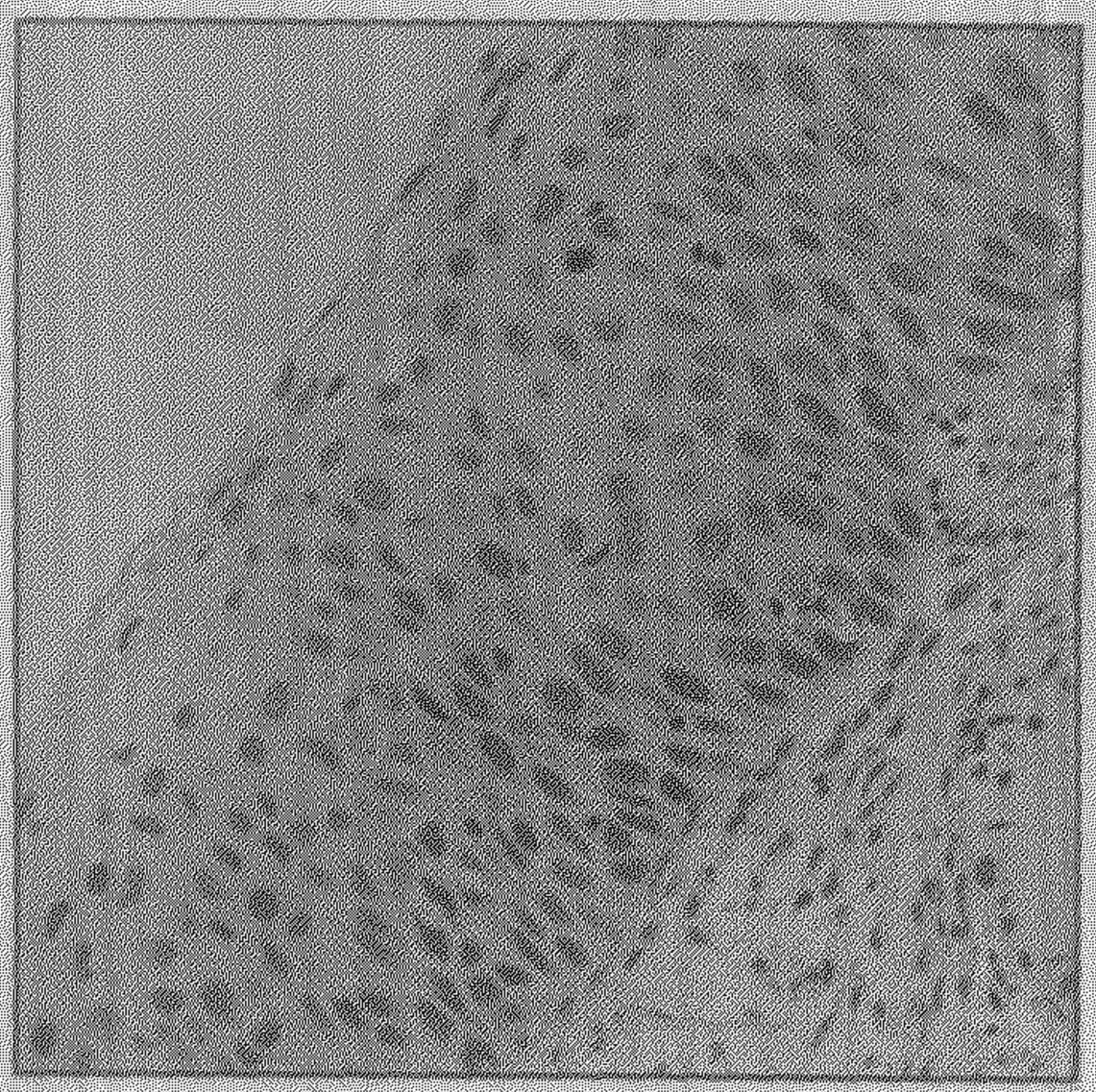


Figure 5.6: Test Image -6

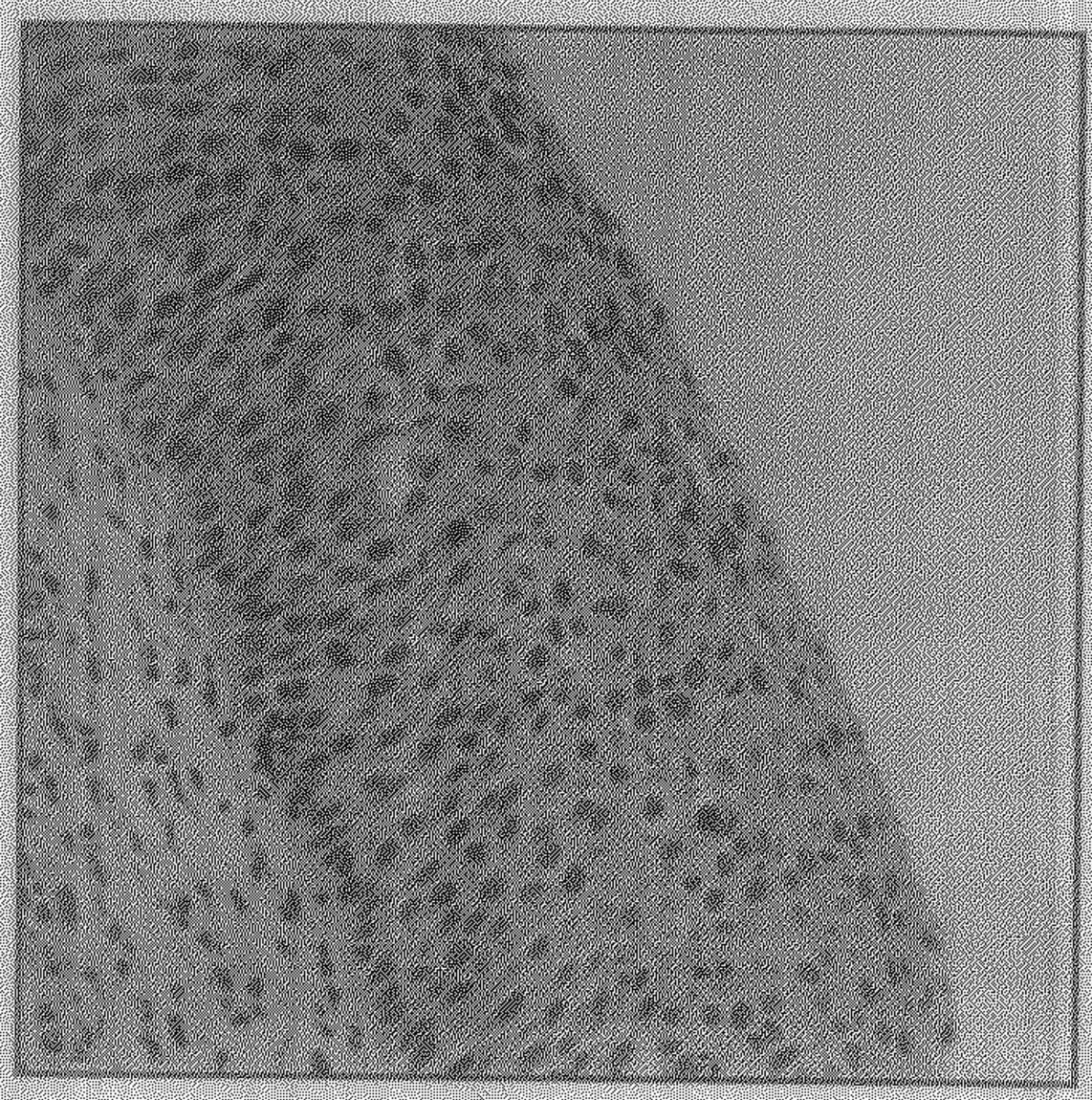


Figure 5.7: Test Image -7

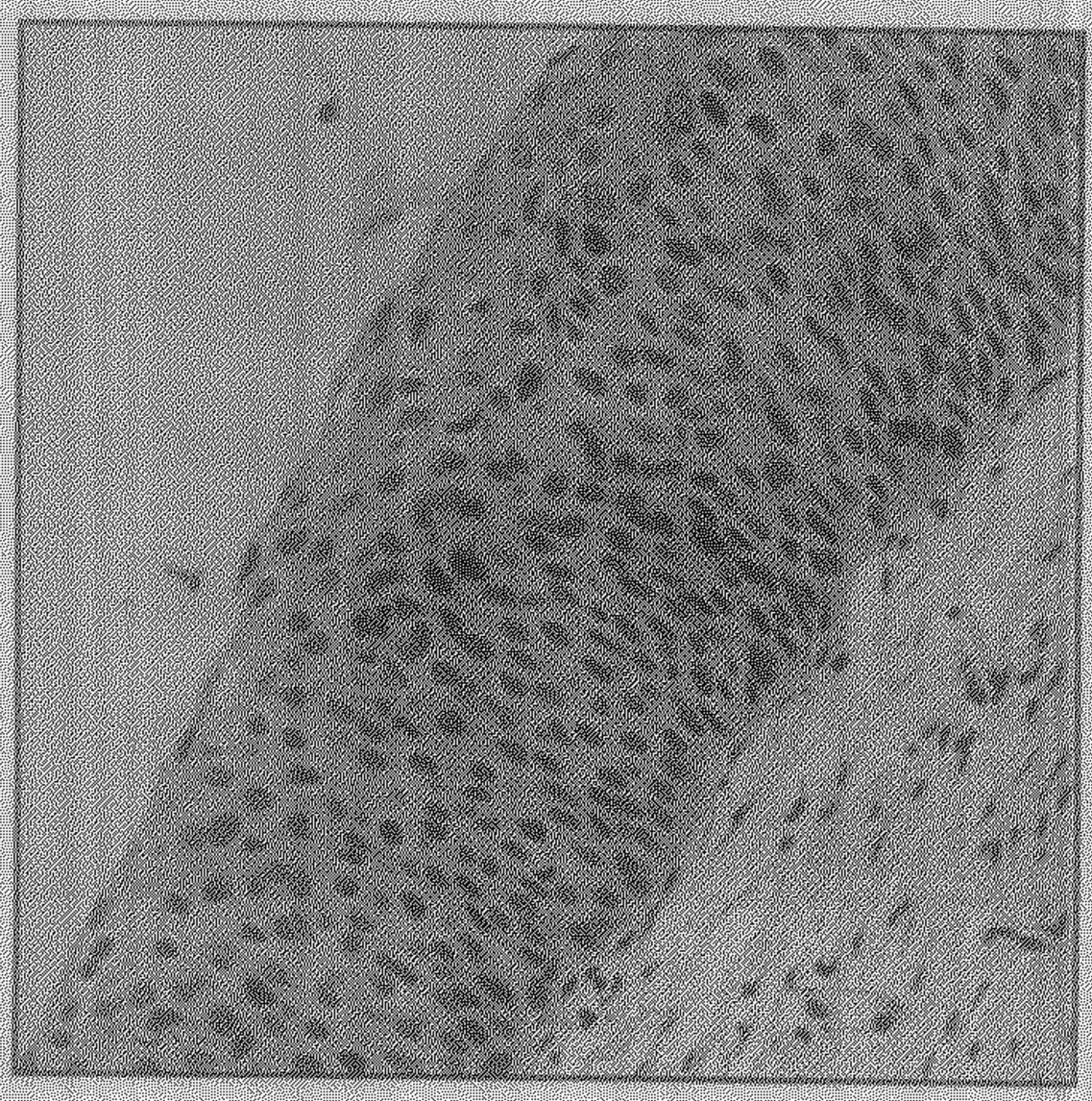
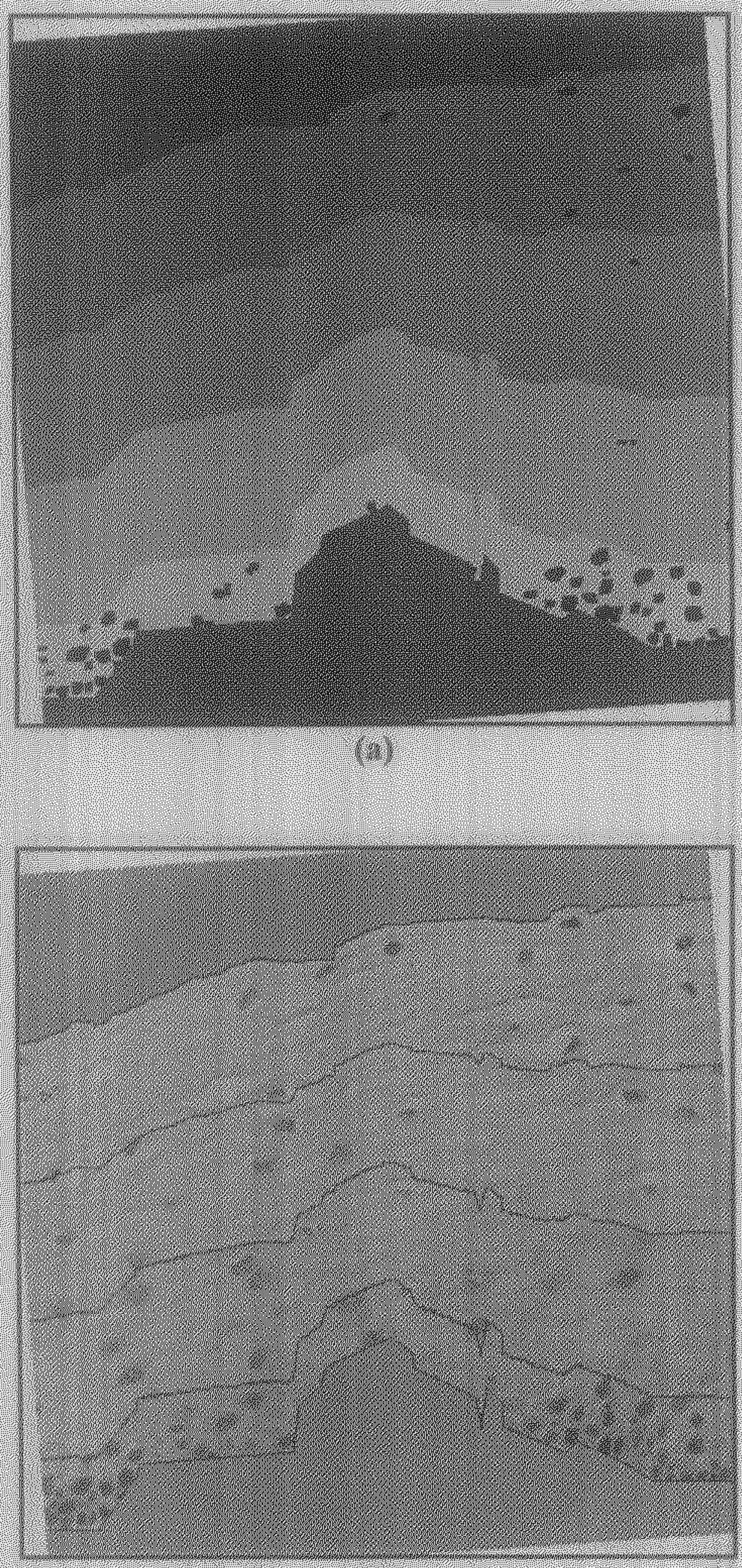


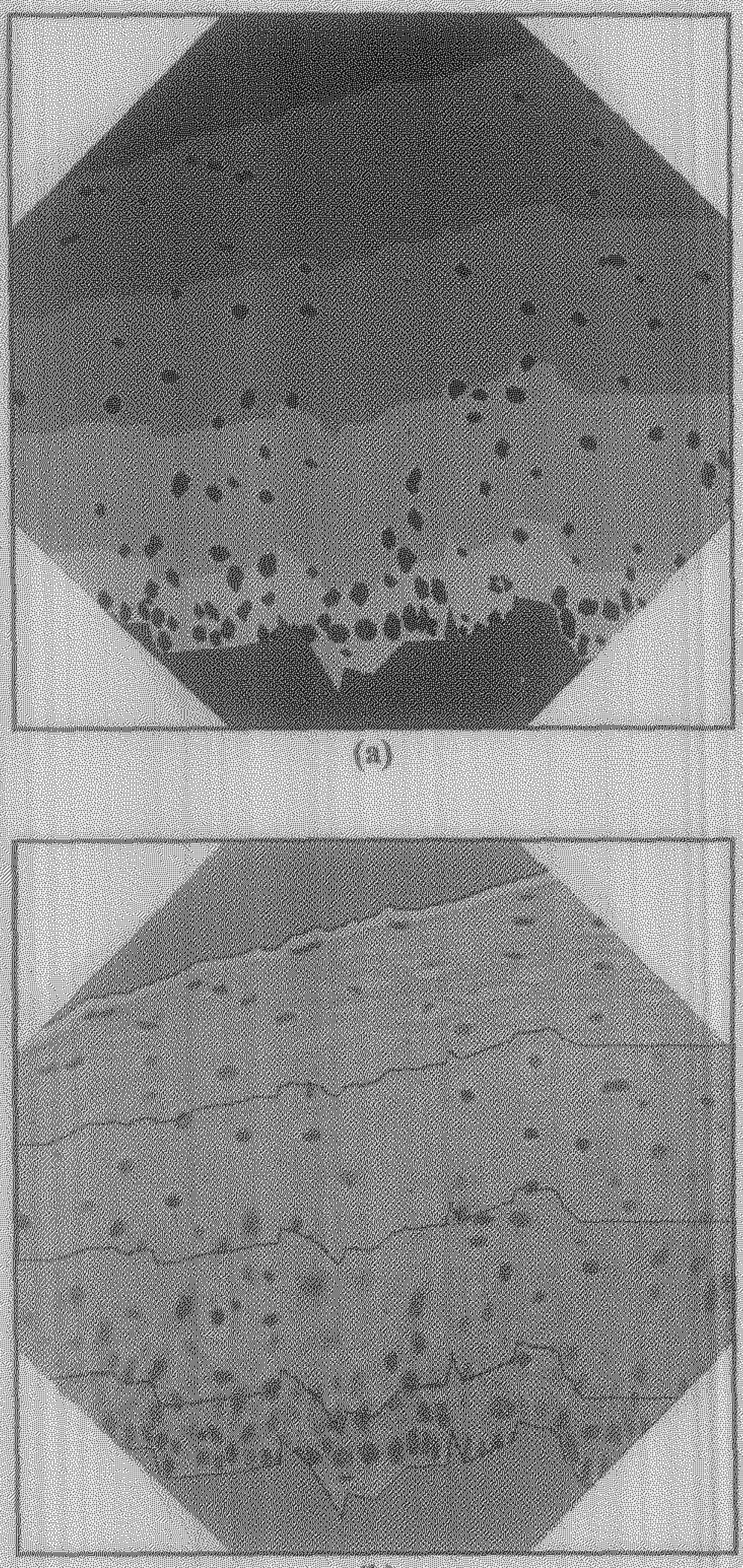
Figure 5.8: Test Image -8



(a)

(b)

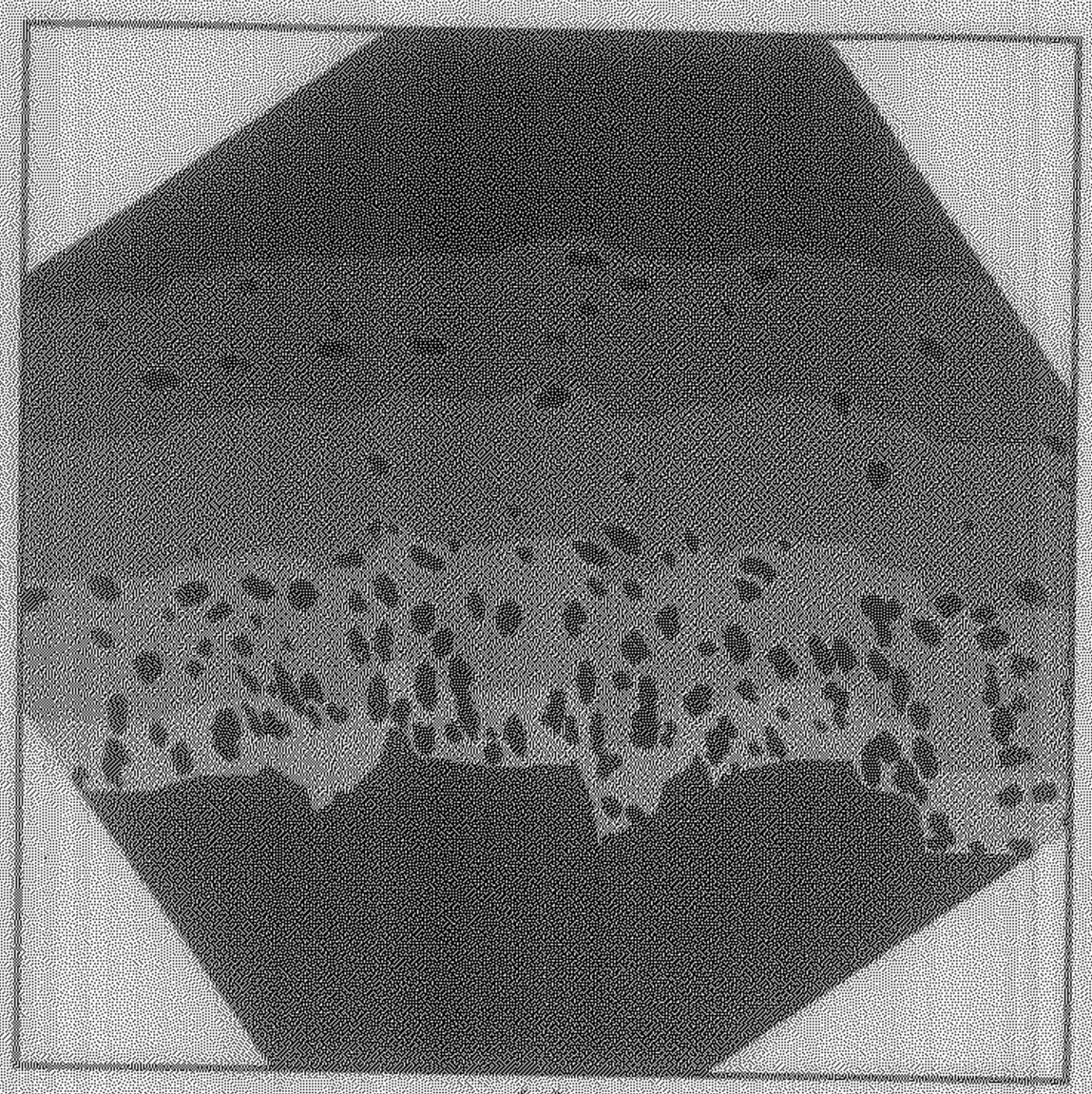
Figure 5.9: Showing four regions for Test Image - 1 (a) with only the nuclei within the regions, (b) with original gray values within the regions



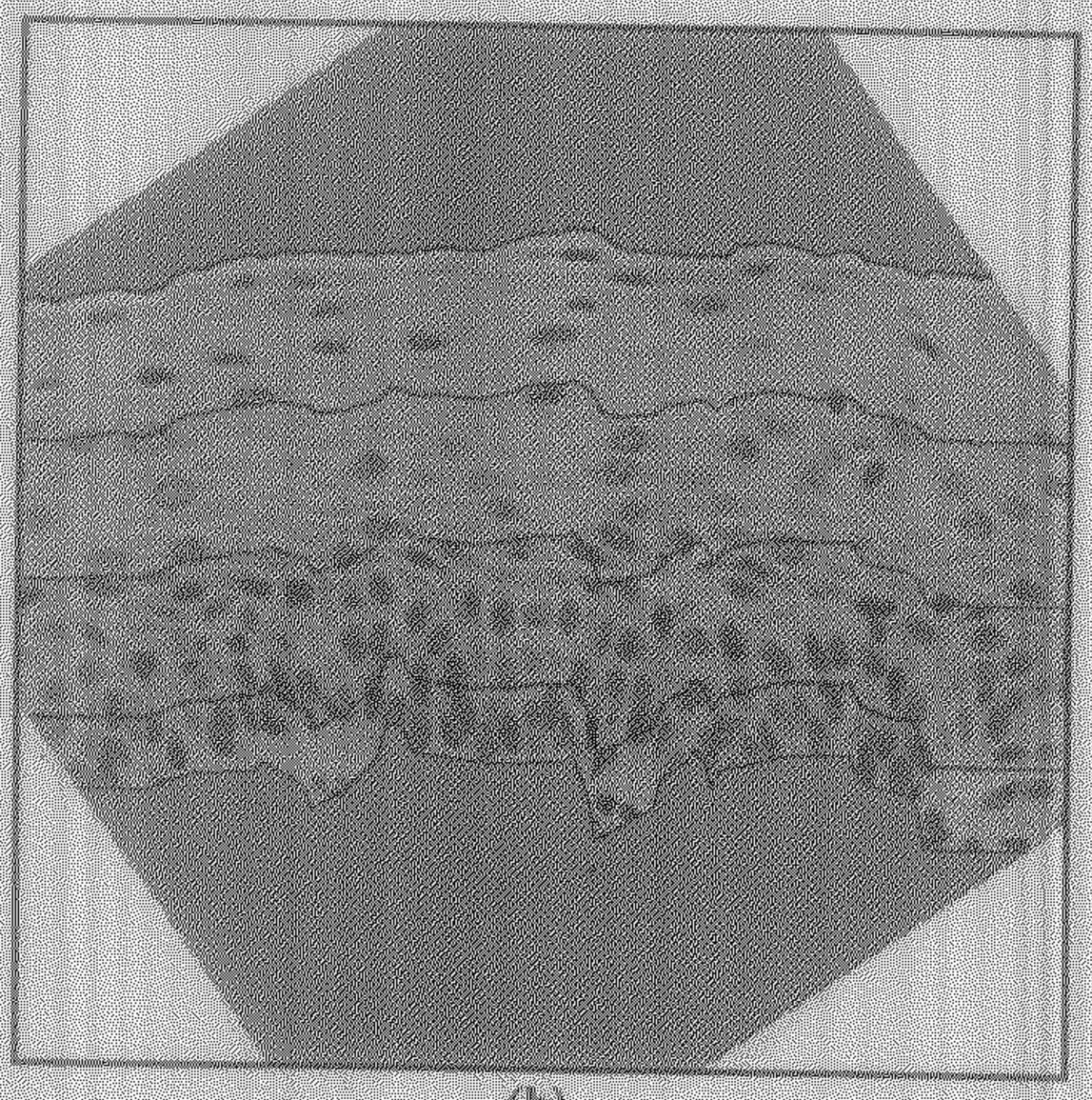
(a)

(b)

Figure 5.10: Showing four regions for Test Image - 2 (a) with only the nuclei within the regions, (b) with original gray values within the regions

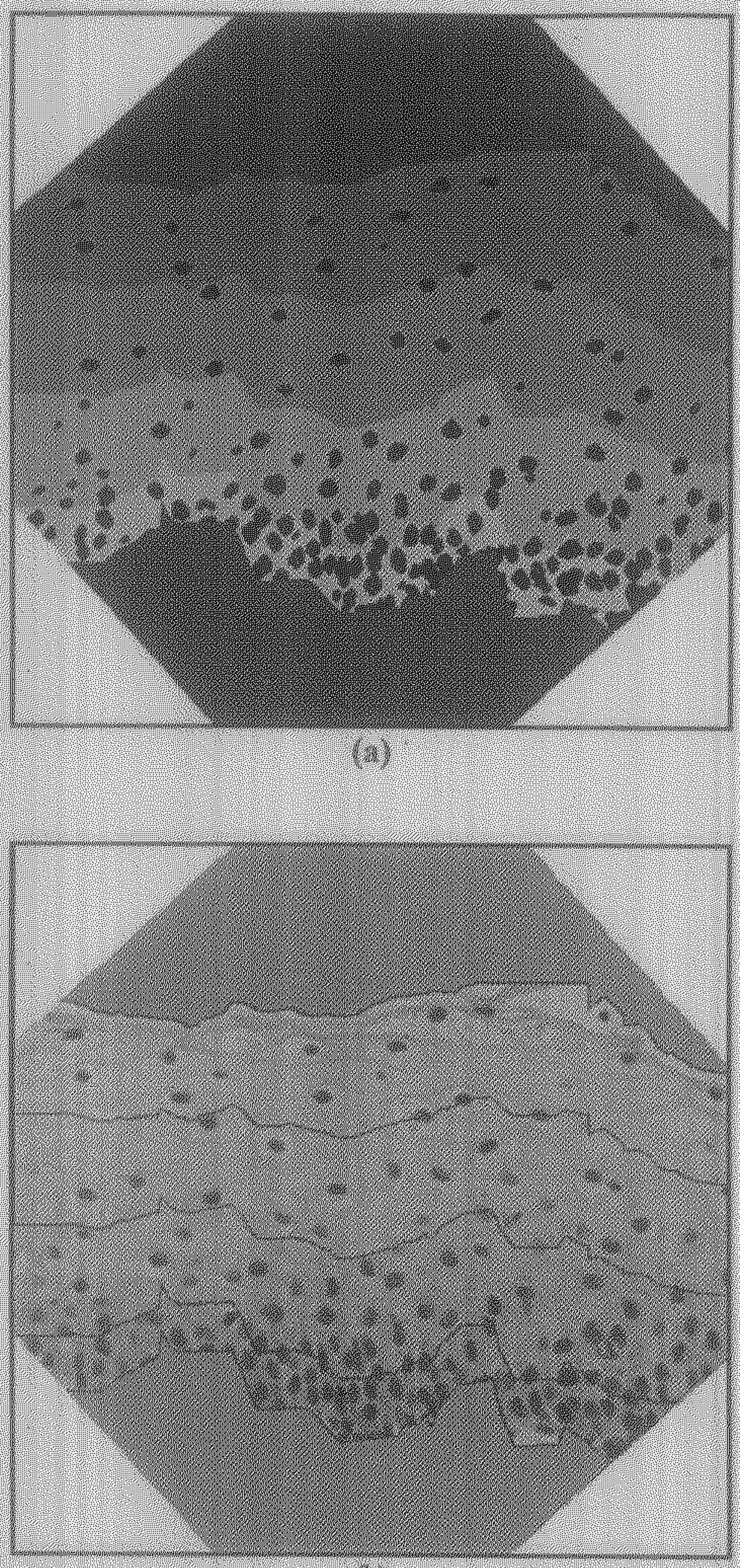


(a)



(b)

Figure 5.11: Showing four regions for Test Image - 3 (a) with only the nuclei within the regions, (b) with original gray values within the regions



(a)

(b)

Figure 5.12: Showing four regions for Test Image - 4 (a) with only the nuclei within the regions, (b) with original gray values within the regions

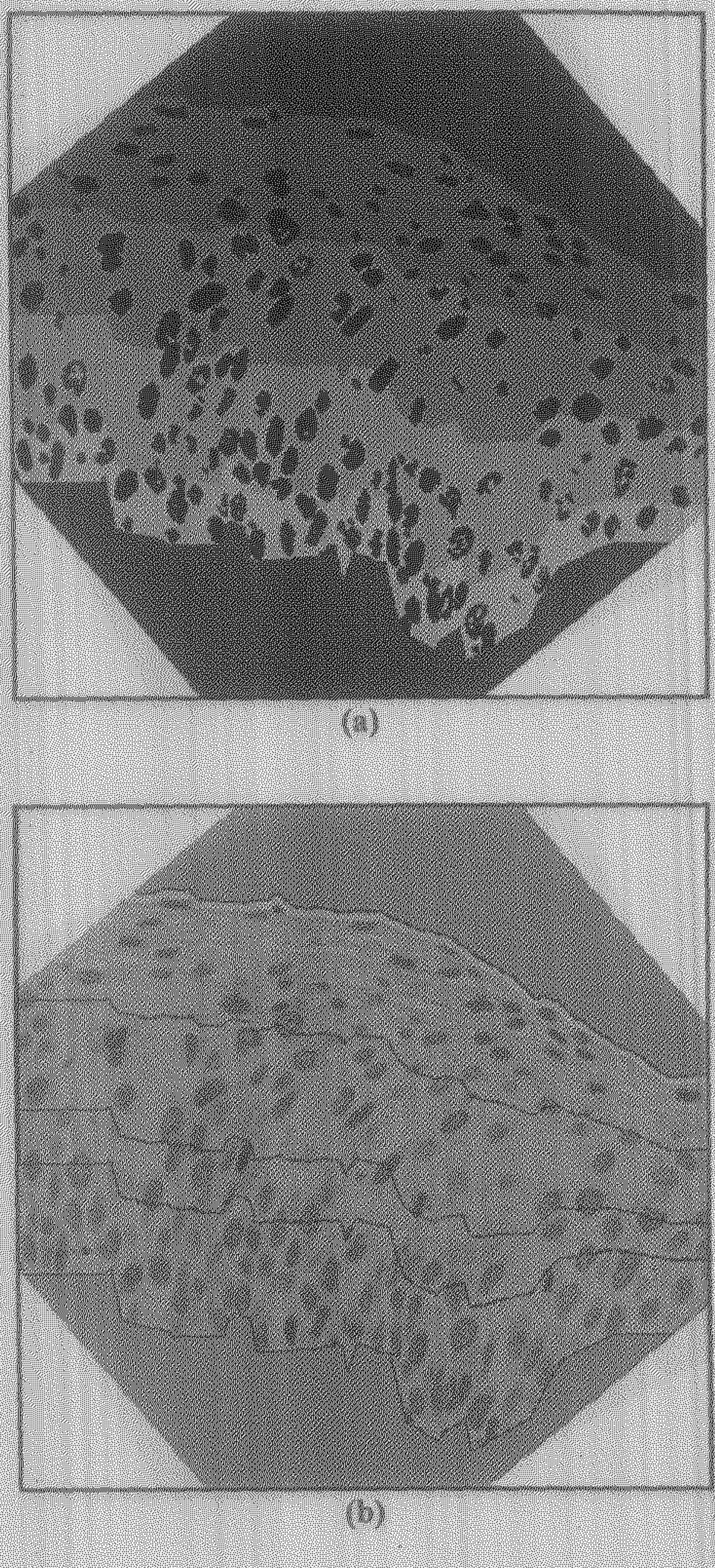
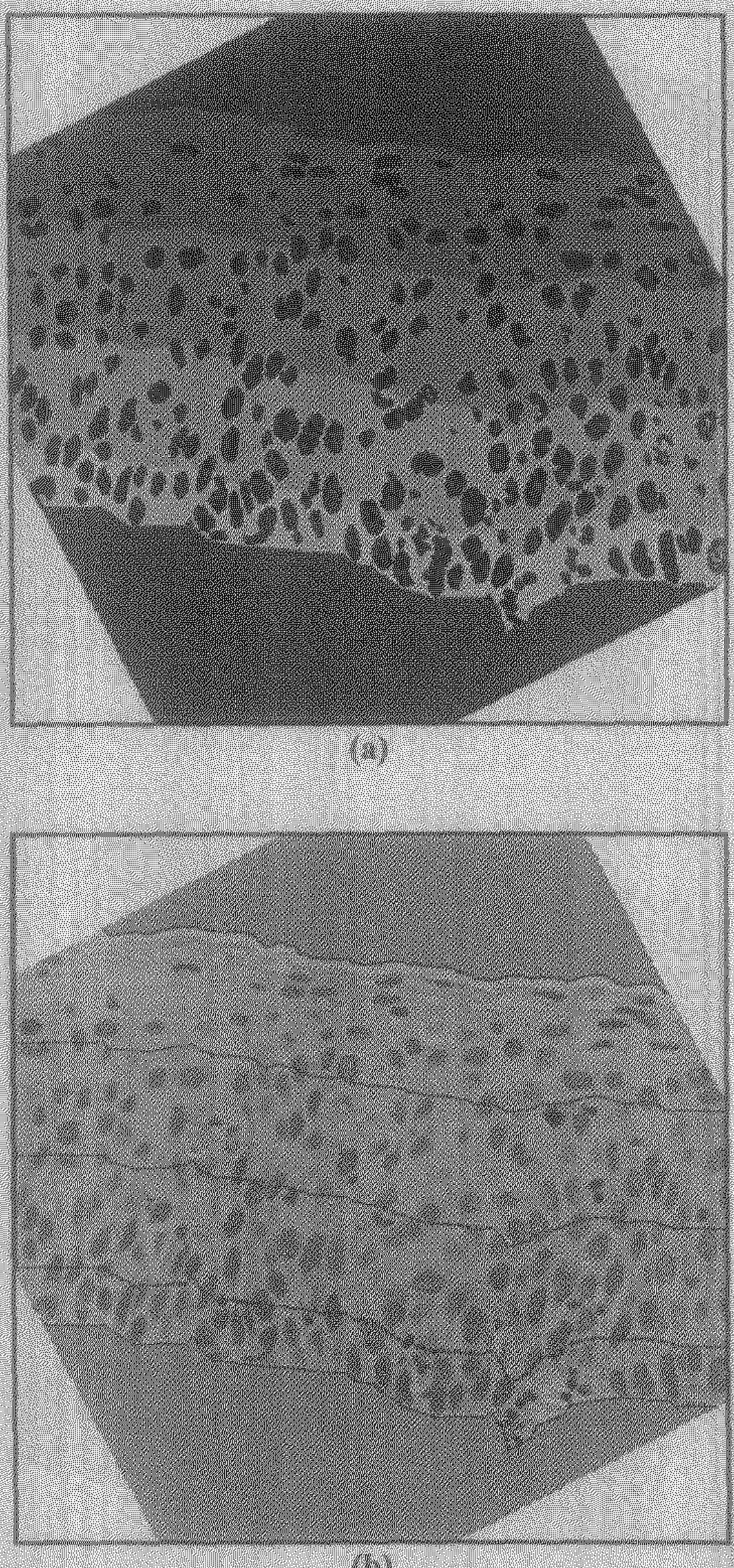


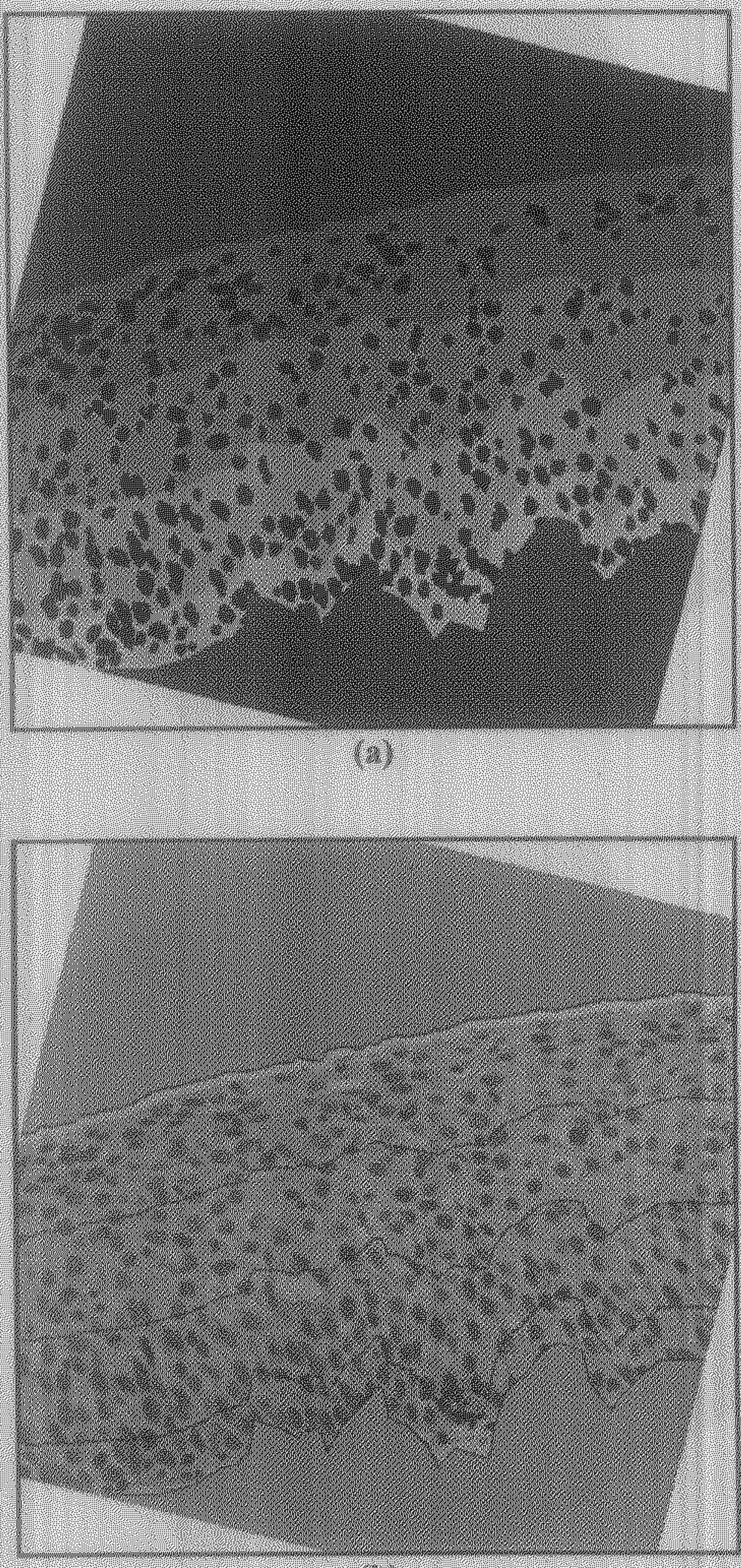
Figure 5.13: Showing four regions for Test Image - 5 (a) with only the nuclei within the regions, (b) with original gray values within the regions



(a)

(b)

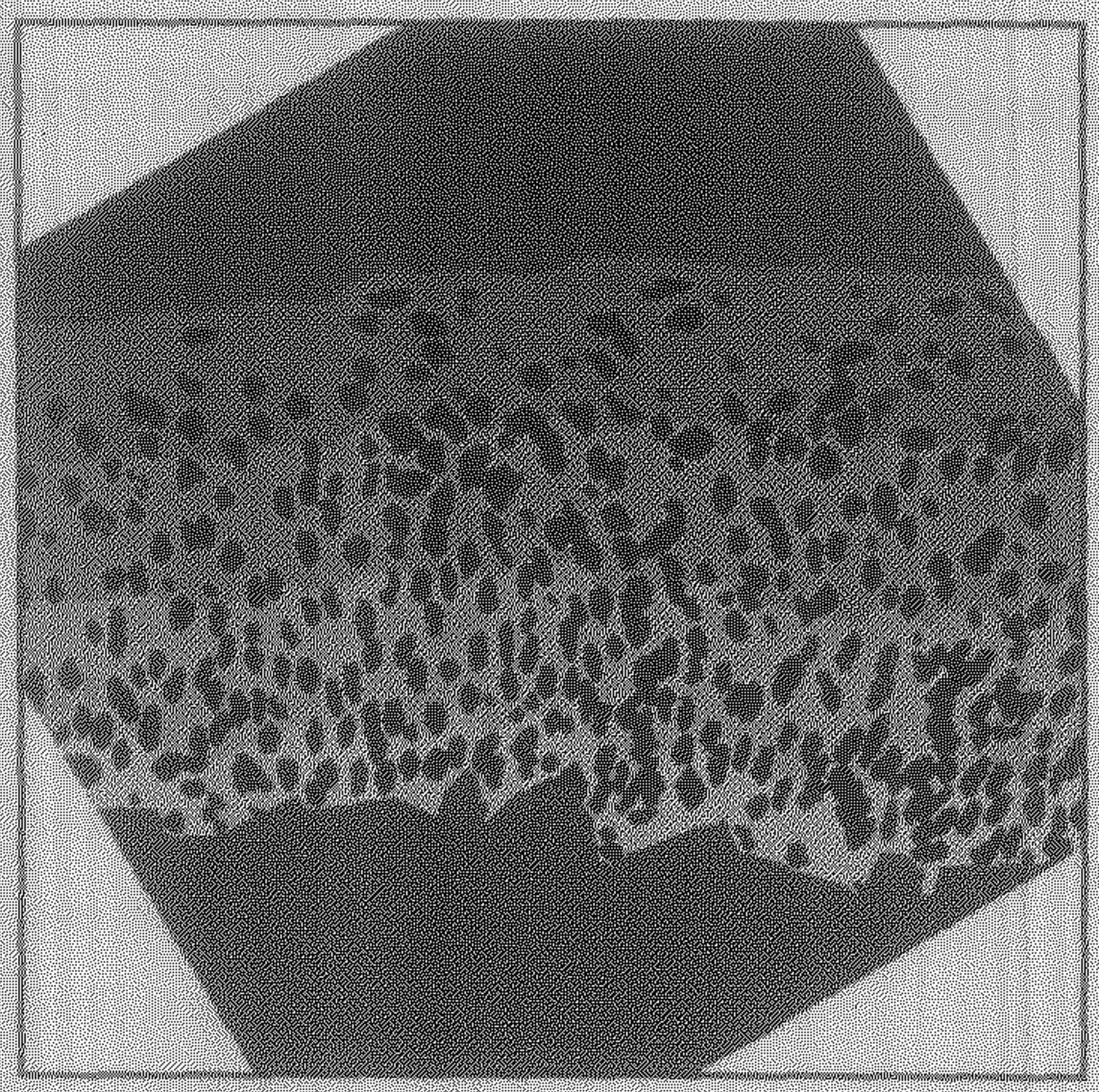
Figure 5.14: Showing four regions for Test Image - 6 (a) with only the nuclei within the regions. (b) with original gray values within the regions



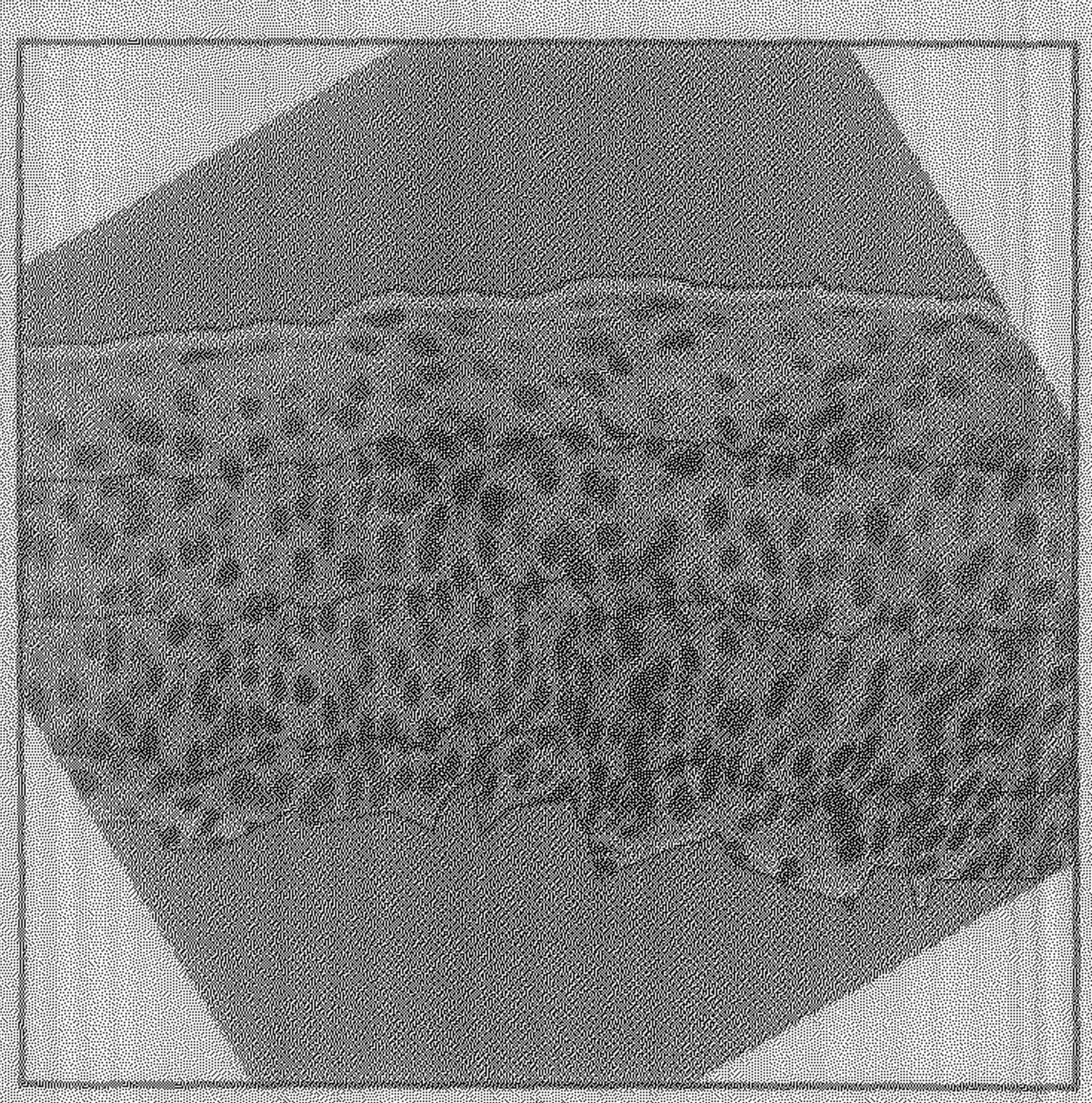
(a)

(b)

Figure 5.15: Showing four regions for Test Image - 7 (a) with only the nuclei within the regions, (b) with original gray values within the regions



(a)



(b)

Figure 5.16: Showing four regions for Test Image - 8 (a) with only the nuclei within the regions, (b) with original gray values within the regions

5.2 Feature Extraction and Classification

After identification of the 4 types of regions, the next important task left is to classify each slide as either normal, CIN-1, CIN-2 or CIN-3. To achieve this, we need to compute some discriminating features for each region. Note that oncologists report their decisions mostly depending on the position of immature cells on respective four regions. Immature cells are characterized by the size of nuclei, relative density of nuclei, orientation of nuclei, gray value variation in each nucleus, etc. Therefore, a possible set of useful features may be as below :

(i) Number of nucleus pixels (F_1),

(ii) Number of non nucleus pixels (F_2),

(iii) Number of (cell) nuclei (F_3),

(iv) Average nucleus size (F_4),

(v) Minimum nucleus size (F_5),

(vi) Maximum nucleus size (F_6),

(vii) Number of horizontally elongated nuclei (F_7),

(viii) Number of vertically elongated nuclei (F_8),

(ix) Composite standard deviation of nuclei (F_9),

(x) Minimum standard deviation of nuclei (F_{10}),

(xi) Maximum standard deviation of nuclei (F_{11}),

(xii) Composite average gray value of the nuclei (F_{12}),

Table 5.1: Feature values for reference image

Regions	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
Superficial	6628	28137	48	133	6	283	19	29	28.05	11.56	41.58	89.30
Intermediate	12321	22315	62	202	4	834	55	7	27.22	2.27	42.65	88.03
Parabasal	12357	21279	78	173	4	1791	71	7	28.32	11.61	42.42	89.26
Basal	7720	9874	32	230	4	1049	29	3	28.89	11.43	43.88	86.84

Table 5.2: Feature values for Test Image - 1

Regions	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
Superficial	376	51397	6	62	35	113	0	6	29.73	22.93	37.81	89.87
Intermediate	27	51925	1	27	27	27	0	1	24.21	24.21	24.21	92.37
Parabasal	54	51374	3	18	15	23	1	2	14.65	10.64	17.19	97.22
Basal	3516	22685	39	93	18	428	7	32	25.54	10.73	35.62	87.36

Table 5.3: Feature values for Test Image - 2

Regions	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
Superficial	647	47982	12	52	24	81	0	12	32.62	14.72	56.91	94.91
Intermediate	1600	55086	21	81	25	129	10	11	29.51	14.29	48.42	99.19
Parabasal	3169	52664	33	93	15	188	23	10	23.86	7.39	49.87	98.13
Basal	5052	17861	39	124	26	382	28	11	24.06	10.35	41.55	96.85

Table 5.4: Feature values for Test Image - 3

Regions	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
Superficial	1145	35987	16	70	19	135	3	13	26.92	12.01	43.45	103.52
Intermediate	1200	36483	17	76	18	168	7	10	22.48	7.76	39.68	100.18
Parabasal	7727	29945	60	129	7	410	37	23	23.29	8.31	39.29	102.29
Basal	4904	13950	33	145	9	535	23	10	21.77	6.99	41.92	104.58

Table 5.5: Feature values for Test Image - 4

Regions	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
Superficial	**1276**	41265	14	87	26	129	1	13	40.42	23.22	59.07	86.09
Intermediate	**1696**	40838	20	86	16	134	3	17	37.61	23.23	50.82	89.29
Parabasal	**6260**	35721	48	123	8	409	28	20	30.45	9.77	41.83	87.67
Basal	**7611**	11236	39	198	29	873	27	12	30.52	10.54	42.58	81.22

Table 5.6: Feature values for Test Image - 5

Regions	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
Superficial	5625	36128	47	123	21	437	8	39	23.57	5.73	56.36	122.15
Intermediate	7822	36662	41	187	5	656	26	15	17.92	1.70	38.48	129.19
Parabasal	9927	34557	56	178	5	594	40	16	15.92	6.39	43.02	127.17
Basal	7354	15378	36	194	4	683	32	4	18.27	6.45	45.59	122.61

Table 5.7: Feature values for Test Image - 6

Regions	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
Superficial	6547	36040	48	141	25	297	14	34	26.55	7.21	47.61	116.13
Intermediate	9843	34868	62	157	23	598	53	9	22.92	4.31	48.54	118.27
Parabasal	13484	31227	64	209	14	816	54	10	21.12	3.15	47.83	116.63
Basal	9330	13326	35	258	26	630	34	1	27.30	9.23	52.25	109.08

Table 5.8: Feature values for Test Image - 7

Regions	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
Superficial	7715	32653	67	115	16	662	30	37	26.58	4.36	39.77	90.75
Intermediate	7911	31963	85	92	24	352	60	25	27.11	6.11	47.14	89.92
Parabasal	11995	27617	84	145	25	540	74	10	25.09	10.04	40.15	89.08
Basal	7353	12276	42	182	41	736	31	11	28.63	8.91	48.39	82.84

Table 5.9: Feature values for Test Image - 8

Regions	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}
Superficial	7405	31209	54	138	4	556	17	37	26.12	4.60	41.30	91.10
Intermediate	13444	25565	66	199	31	802	53	13	27.23	9.48	44.37	88.33
Parabasal	16442	22401	69	266	23	2435	66	3	28.71	8.54	43.57	87.00
Basal	8489	10868	33	206	26	643	27	6	31.76	13.49	52.12	83.40

A nucleus is considered to belong in the region where its center lies. Average gray value and standard deviation of gray values for each nucleus are computed based on the original gray value in the image. The mean of this averages and mean of standard deviations over respective regions are the composite average gray value and composite standard deviation, respectively. In case, height of a nucleus is more than its width, the nucleus is considered vertically elongated, otherwise, it is considered horizontally elongated. The above features are calculated for each of the regions separately.

The Table 5.1 provides the values of the 12 features F_1, F_2, ..., F_{12} for the four regions computed on the image in Fig.4.2. Computed feature values for the eight test images [figures 5.1 - 5.8] are included in tables 5.2 - 5.9.

Analysis of these tables reveals that the ratio of F_1 to F_2 significantly varies with the status of the slide images. However due to lack of time, a detail analysis of the features has not been done.

For each image, we get 48 features (12 for each of the four regions). Let these features be denoted by a vector $X = (x_1, ..., x_{48}) \in R^{48}$. For each such image we also have a label vector, say, $e_i = (e_{i1}, ..., e_{i4})^T \in R^4$ where, $0 \leq e_{ij} \leq 1$, $\forall\ j = 1, ..., 4$, e_{ij} denotes the degree to which the slide falls in the jth category ($j = 1$ means normal, $j = 2$ means CIN-1, $j = 3$ means CIN-2 and $j = 4$ means CIN-3). Note that normally, for most of the slides $e_{ik} = 1$ for some k and $e_{ij} = 0$ for $j \neq k$; However for some distinguished cases, fuzzy label vectors may be assigned by experts also. Let there be N such slides, thus we will get the training set (X, Y), where $X = x_i, ..., x_N \subset R^{48}$ and $Y = e_i, ..., e_N \subset R^4$, and e_i is the label vector of x_i characterizing the position of slide.

Using this training data (X, Y) several classifier may be designed. For example, we can use k-nearest neighbor classifier [5–7], multilayer perceptron network [24, 25, 26], fuzzy rule based classifier [27, 28] etc.

Chapter 6

Conclusions

In the present investigation, an attempt is made towards finding suitable methodologies to identify pre-cancerous changes in the tissues of cervix. These methodologies are based on pattern recognition and image processing techniques. Since only the portion between the basal and superficial membrane is required to analyzed for the said purpose, a major part of our investigation is concerned with identification of the two membranes. Finally, some discriminating features are computed which may be useful to classify the image as either normal, CIN-1, CIN-2 or CIN-3.

For decomposing the overlapping nuclei, a morphological splitting algorithm is formulated. We have proposed a method which can rotate any slide image such that the basal and superficial membranes become almost horizontal and basal membrane is placed below the other one.

The output of FCM algorithm with the texture features shading and prominence is found to be effective for identifying the superficial membrane. We have also shown that FCM with gray values, average gray values and s.d. is able to extract most of the nuclei. We feel that FCM algorithm with various other combination of textural features may be useful.

The discriminating features, we have computed, are based on the size, relative density, orientation, gray value variation etc., of nuclei to identify immature cells. The ratio of nucleus area with its surrounding cytoplasm area may be an important feature to characterize an immature cells. This needs further investigation. Our system also requires several parameters whose choices are important. Some

guidelines in this regard would be useful which we plan to do in near future.

The proposed system is developed using the on slide images supplied by GSF foundation, Munich, Germany. It is necessary to study the effectiveness of the proposed methodologies on slide images, collected from Indian laboratories. We have mentioned a few possible approaches to design a classification system. An attempt is being made to design a classifier for the proposed system.

Bibliography

[1] H. M. Shingleton, R. L. Patrick, W. W. Johnston and R. A. Smith, "The current status of the Papanicolaou smear," *CA: A Cancer J. Clin.*, vol. 45, pp. 305–320, 1995.

[2] L. A. Brinton : "Epidemiology of cervical cancer- overview, The epidemiology of cervical cancer and human papilloma virus," IARC publication no. 119, pp. 4–23, 1992.

[3] M. S. Piver, R. E. Hempling, and K. A. Craig, "Neoplasms of the cervix", in *Cancer Medicine*, vol. 2, J.F.Holland, F.Frei, R.C.Bast, D.W.Kufe, D.L.Mortan and R.R.Weichselbaurns (Eds.), Lea & Febiger, Philadelphia, 1993.

[4] A. K. Ghosh, *Text Book of Gynaecology*, Current Books International, Calcutta, 1982.

[5] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, New York: Wiley Interscience, 1973.

[6] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles*, Reading, MA: Addison-Wesley, 1974.

[7] P. A. Devijver and J. Kittler, *Pattern Recognition : A Statistical Approach*, London: Prentice-Hall, 1982.

[8] S. K. Pal and D. Dutta Majumder, *Fuzzy Mathematical Approach to Pattern Recognition*, New York: John Wiley & Sons, 1986.

[9] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms* New York: Plenum Press, 1981.

[10] J. C. Bezdek and S. K. Pal (eds.), *Fuzzy Models for Pattern Recognition: Methods that Search for Structures in Data*, New York: IEEE Press, 1992.

[11] L. A. Zadeh, "Fuzzy sets," *Information & Control*, vol. 8, no. 3, pp. 338–353, 1965.

[12] J. C. Bezdek, "Cluster validity with fuzzy sets," *J. Cybern.*, vol. 3, no. 3, pp. 58–72, 1974.

[13] A. Rosenfeld and A. C. Kak, *Digital Picture Processing*, New York: Academic Press, 1982.

[14] J. Serra, *Image Analysis and Mathematical Morphology*, London: Academic Press, 1982.

[15] R. C. Gonzalez and P. Wintz, *Digital Image Processing*, Reading, MA: Addison-Wesley, 1987.

[16] R. M. Harralick and G. Shapiro, *Computer and Robot Vision*, New York: Addison-Wesely, 1991.

[17] K. S. Fu and J. K. Mui, "A survey on image segmentation", *Patt. Recog.*, vol. 13, pp. 3–16, 1981.

[18] N. R. Pal and S. K. Pal, "A review on image segmentation techniques", *Patt. Recog.*, vol. 26, no. 9, pp. 1277–1294, 1993.

[19] N. R. Pal and S. K. Pal, "Entropic thresholding", *Signal Process.*, vol. 16, no. 2, pp. 97–108, 1989.

[20] J. N. Kapur, P. K. Sahoo and A. K. C. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram", *Comput. Graph. Vision Image Process.*, vol. 29, pp. 273–285, 1985.

[21] C. E. Shannon, "A mathematical theory of communication", *Bell System Tech. Journ.*, vol. 27, pp. 379–423;623–656, 1948.

[22] S. Guiasu, *Information Theory with Applications*, McGraw-hill Int. Book Company, New York, 1977.

[23] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural features for image classification," *IEEE Trans. Systems, Man Cyberns.*, vol. 3, pp. 610–621, 1973.

[24] D. E. Rumelhart, J. McClelland, *et al.*, *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*, Vol. 1, Cambridge, MA: MIT Press, 1986.

[25] Y. H. Pao, *Adaptive Pattern Recognition and Neural Networks*, New York: Addison-Wesley, 1989.

[26] R. P. Lippmann, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, pp. 4–22, April 1987.

[27] S. L. Chiu, "Fuzzy model identification based on cluster estimation," *J. Intelligent and Fuzzy Systems*, vol. 2, pp. 267–278, 1994.

[28] S. L. Chiu, "Extracting fuzzy rules for pattern classification by cluster estimation," in *Proc. Sixth Fuzzy System Assoc. World Congress (IFSA '95)*, Sao Paulo, Brazil, pp. 1–4, 1995.