To the Graduate Council:

I am submitting herewith a thesis written by Nenad Vujovic entitled "Document Authentication." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Electrical Engineering.

Dragana Brzakovic, Major Professor

We have read this thesis
and recommend its acceptance:

M. P. Abt

Yuh-er Kuo

Accepted for the Council:

Associate Vice Chancellor
and Dean of The Graduate School

DOCUMENT AUTHENTICATION

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Nenad Vujovic

August 1992

Dedication

This thesis is dedicated to my parents, Svetomir and Slavka, and to my sister, Vera.

Acknowledgements

The author wishes to express his profound appreciation to his major advisor, Dr. Dragana Brzakovic, for the privilege of working with her, as well as the invaluable support, advice, and unwavering patience and encouragement that she has given to him throughout his masters program. Many thanks are due to Dr. Abidi, and Dr. Kuo for serving on author's thesis committee, reviewing this thesis and teaching him so much throughout his graduate study.

Special thanks are extended to the University of Tennessee, Knoxville, and the Department of Electrical and Computer Engineering for providing an excellent graduate study program and research environment and equipment. The author is also grateful to Mike Ramsey for supporting with a graduate research assistantship through this research and providing many challenging problems.

The author is deeply indebted to his friends Predrag Bakic and Milorad Neskovic for their constant support throughout his undergraduate and graduate studies. Finally, and most importantly, the author would like to thank his parents for their sturdy motivation and strong support.

Abstract

This thesis deals with document authentication. The test document is compared with documents in the database in order to determine a possible match. The database consists of a large number of different documents containing fibers randomly distributed over noisy backgrounds. The test document may be translated and rotated relative to documents in the database. Also, some fibers may be missing in the test document relative to those in the database. In this thesis a new approach to document authentication based on distances between fiber end points is introduced. The algorithm has been tested on more than a thousand different documents. The thesis analyzes the probability of uniqueness of documents represented by fibers as well as the probability of uniqueness of documents represented by fiber end points. To overcome difficulties caused by the large database, assigning a unique signature to documents is considered in the thesis, too. Three different ways of assigning signatures to documents are considered, and all of them are based on the concept of distances between fiber end points.

TABLE OF CONTENTS

CHAPTER PAGE

1. Introduction 1

2. Synthetic Image Generation 7

2..1 Background generation 7

2..2 Fiber generation 10

3. Fiber Extraction **24**

3.1 Dynamic range compression 25

3.2 Gradient as an edge detector 29

3.3 Detection of the end points of fibers 36

4. Document Authentication **40**

4.1 Database generation 42

4.1.1 File generation 42

4.2 Matching 44

4.3 Point matching 45

4.3.1 Determining the first point match 47

4.3.2 Determining the second point match 49

4.3.3 Determining the third point match 50

4.4 Image matching 52

CHAPTER PAGE

5. Probabilities of Uniqueness **55**

5.1 Case 1: Line segments 55

5.1.1 Fiber case 55

5.1.2 End point case 58

5.2 Case 2: Fibers of arbitrary shapes 60

5.2.1 Fiber case 60

5.2.2 End point case 60

6. An Alternative Approach to Document Authentication . . . **65**

6.1 An approach based on histogram of distances 66

6.2 An approach based on an ascending order of distances 73

6.3 An approach based on distances and phase angles 78

7. Conclusions **92**

BIBLIOGRAPHY **94**

Vita **97**

LIST OF FIGURES

FIGURE		PAGE
1.1	An example of the original documents: (a) entire document (b) an enlarged part of the document from (a) (from ORNL)	3
2.1	Some examples of backgrounds (image size is 256×256 and minimum and maximum intensities are 0 and 200, respectively): (a) $f = 0.2$, $n = 8$; (b) $f = 0.8$, $n = 8$; (c) $f = 1.0$, $n = 2$; (d) $f = 1.0$, $n = 15$.	11
2.2	(a) A fiber simulated by a circle arc: r, l, and α are random variables normally distributed. (b) A fiber simulated by a line segment: both l and α are random variables normally distributed.	16
2.3	An example of fiber images: (a) one pixel fiber widths and (b) five pixel fiber widths.	17
2.4	Typical examples of synthetic images.	19
3.1	An original document.	26
3.2	(a) Histogram of the background of the image in Figure 3.1. (b) Histogram of fibers of the image in Figure 3.1.	27
3.3	Result of the gradient technique applied to the original image in Figure 3.1.	30
3.4	Preprocessed image.	31
3.5	Gradient image of the preprocessed image.	32

3.6 (a) Mask used to compute G_x at (x, y). (b) Mask used to compute G_y at (x, y). 34

3.7 End points of fibers of a document superimposed on the document itself. End points were found by the algorithm for end point detection. 39

4.1 An image divided into four quadrants. 43

5.1 Uniqueness as a function of fiber density for fibers as line segments. 59

5.2 Uniqueness as a function of fiber density for documents consisting of fibers of arbitrary shapes. 61

5.3 Uniqueness of the end point case: fibers are of arbitrary shapes. . 64

6.1 Histograms of distances of two different images. 67

6.2 Histograms of an image with (a) one missing fiber, (b) two missing fibers, and (c) twenty missing fibers. 70

6.3 Ascending order of distances of an arbitrary document. 74

6.4 Ascending order of distances of a document with (a) one, (b) two, and (c) twenty fibers missing (dashed lines); ascending order of distances of the original document is shown by solid lines. 75

6.5 Ascending order of distances of two different documents. 79

6.6 Graphs $r(\theta)$ for two different documents. 80

6.7 Histograms of graphs $r(\theta)$ from Figure 6.6. 82

6.8 Difference between histograms (a) and (b) from Figure 6.7. . . . 84

6.9 Graph $r(\theta)$ from Figure 6.6 (a) when (a) one, (b) two, and (c) twenty fibers are missing. 86

6.10 Histograms from Figure 6.9. 89

CHAPTER 1

Introduction

This thesis deals with document authentication, an algorithm that establishes whether or not a test document belongs to a prestored database. Generally, the database contains a large number of documents. A test document is subject to three types of deformations:

(1) the test image may be translated with respect to its prestored version

(2) the test image may be rotated with respect to its prestored version

(3) the test image may contain a subset of fibers with respect to its prestored version; i.e., some fibers may be physically destroyed at a later time.

Maximum values for translation, rotation, and missing fibers are prespecified.

The authentication problem is closest to the problem known in the literature as *registration* or *correspondence* ([2], [3], [5], [6], [13], [14], [15]). The registration problem is usually defined as follows. Given an original document and a test document (both with prespecified features), the objective is to determine whether or not the two documents match each other. Sometimes, instead of one original document, a sequence of originals may be considered (for example, a set of images taken from an aircraft or a satellite). The images in a sequence differ mostly in translation and rotation. Features that are considered for matching may be points (control points or tie-points) or patterns that are prespecified. So,

the main objective of the registration problem is to precisely align two images (or sequence of images) prior to a point-to-point comparison of the images.

The document authentication considered in this thesis differs from the correspondence problem in the following ways:

- there are no control points in the documents
- a database contains a large number of different documents.

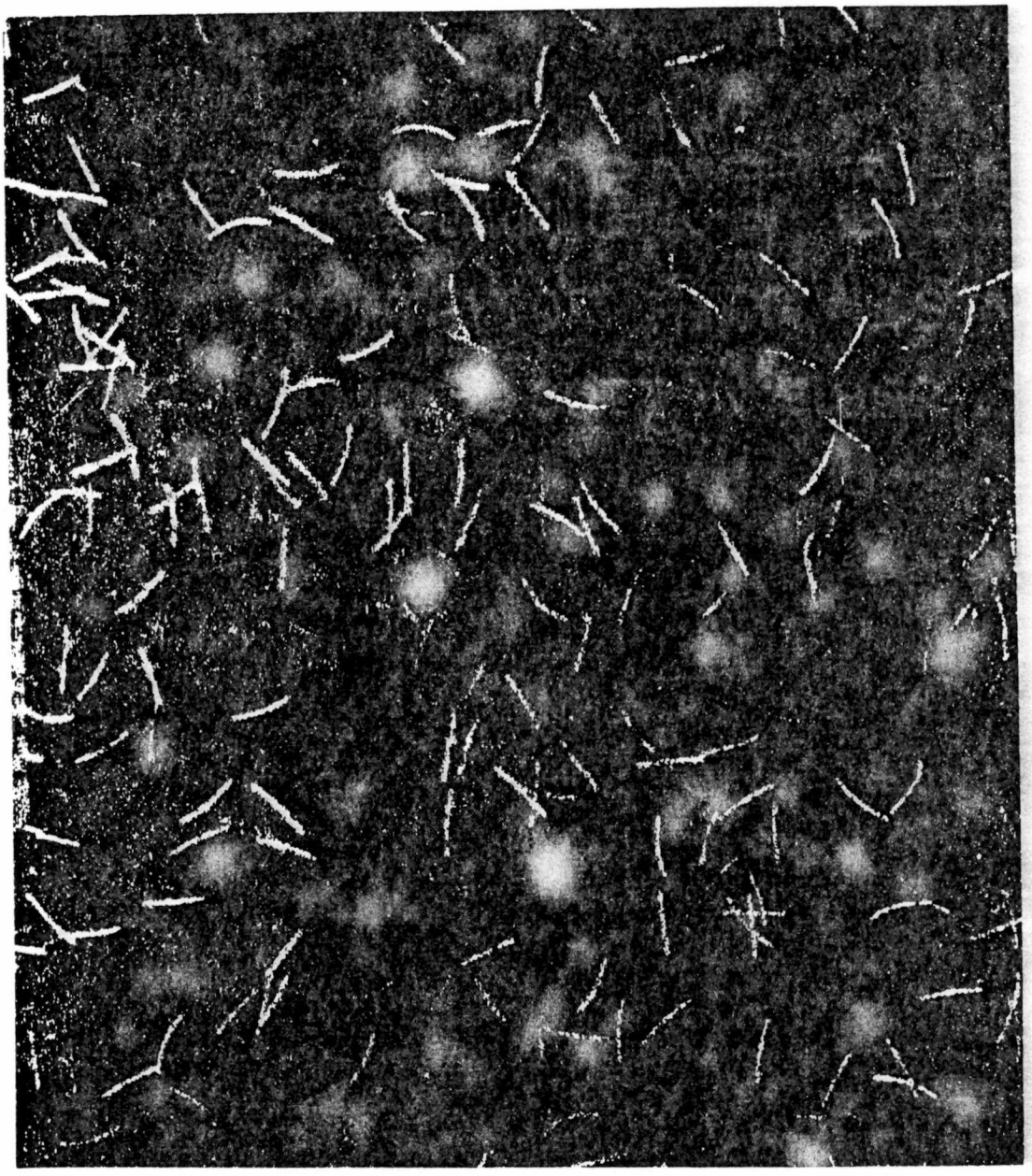
A new approach to the correspondence problem under these assumptions is developed. In addition, alternative approaches that do not require database or arbitrary database searches are considered.

An example of the documents considered in this thesis is given in Figure 1.1. Fibers carry information stored in a document on which fibers are randomly distributed. In general, fibers are of arbitrary shapes and they may overlap.

In Chapter 2, the simulation of real documents is described. Because a document consists of a background and fibers, simulation of backgrounds and simulation of fibers are considered separately. The simulated image is then obtained by superimposing the simulated fibers onto the simulated background.

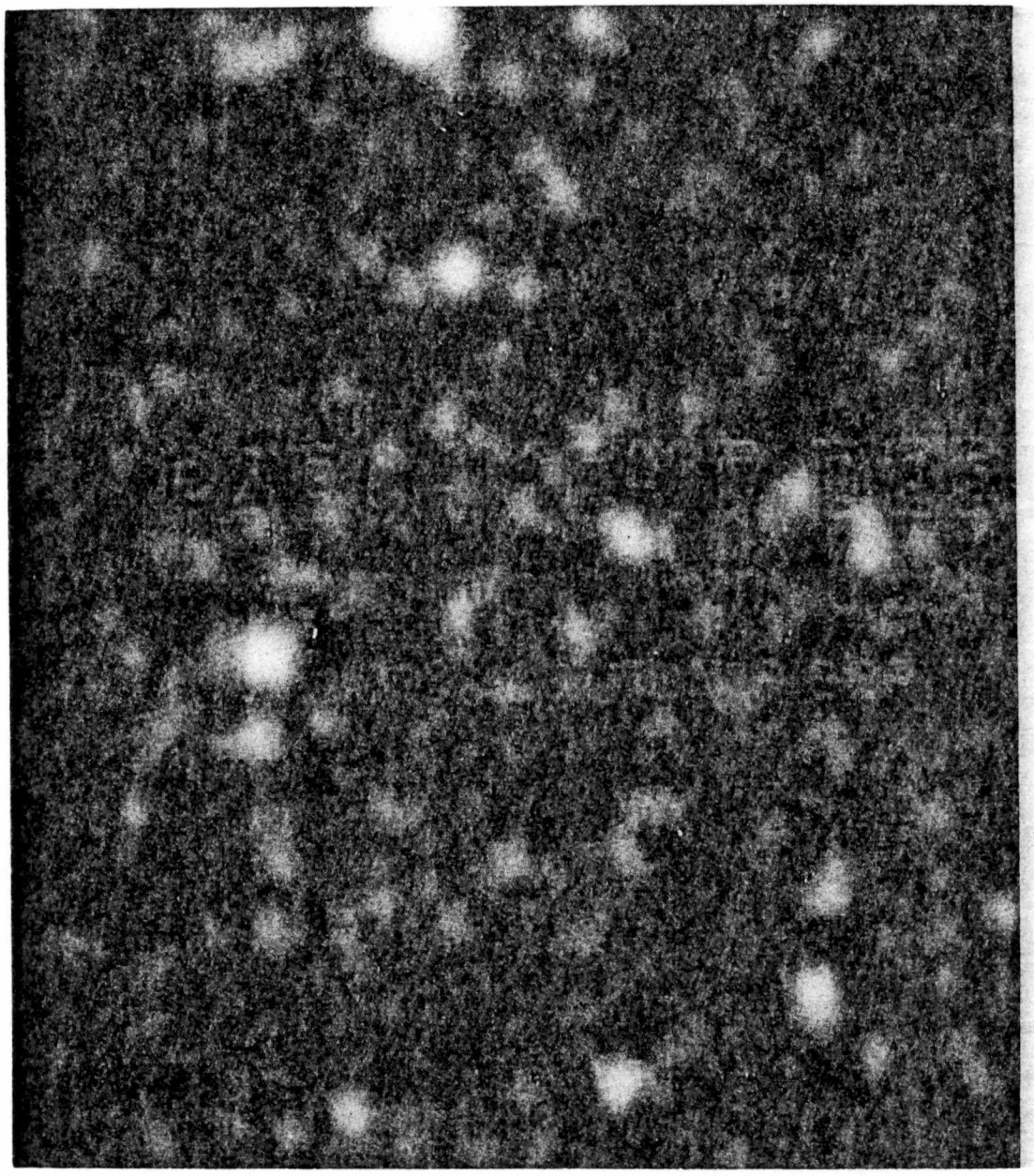
In Chapter 3, fiber extraction from noisy backgrounds of documents is described. There are two steps that are used for efficient fiber extraction: dynamic range compression of the backgrounds and edge detection using one of the gradient techniques. Resulting image is an image with well distinguished fibers and a very small amount of noise.

Chapter 4 describes the authentication procedure. Here, we deal with binary



(a)

Figure 1.1: An example of the original documents: (a) entire document (b) an enlarged part of the document from (a) (from ORNL)



(b)

Figure 1.1: Continued.

images with properly extracted fibers from noisy backgrounds. Those images are referred to as fiber images. The procedure is divided into four steps:

(1) detecting fiber end points in fiber images of documents

(2) establishing if any three points in the test image match three points in any of the database images (end point images are considered)

(3) establishing if the remaining points in the test image match points in the database image chosen in step (2)

(4) establishing if fiber images of the test document and a document chosen in the step (3) match. Document authentication assumes translation, rotation, and missing fibers.

In Chapter 5, the probability of uniqueness of documents is analyzed. The purpose of this calculation is to estimate the number of distinguishable documents represented in the digital form. The uniqueness of fiber documents is considered, taking into account line segment fibers first and then fibers of arbitrary shapes. Next, the effects of data reduction when using only fiber end points is considered.

In Chapter 6, an alternative to generating an arbitrary database is considered. Potentially, there are two different approaches:

(1) to generate an organized database where the physical address of an image corresponds to its contents

(2) to assign a code directly to the document and verify that the code correspond to the document itself. In both cases the problem reduces to generating a unique code corresponding again to its content in the image. This code can

be used as either the physical address of the document, or it can be imprinted on the document itself. In this thesis, three approaches to generating a code are considered. The first approach is based on the histogram of distances between fiber end points. The second is based on an ascending order of all distances between all points in the end point image. The third approach considers distances of all points in the end point image to the reference point as a function of phase angles θ. The reference point is prespecified.

CHAPTER 2

Synthetic Image Generation

This section describes synthetic document generation used for developing the authentication procedure. The generation of backgrounds and the generation of fibers are considered separately. Each document is represented as a digital image and hereafter is referred to as an image.

Computer simulation offers an inexpensive way of generating a large number of documents. Furthermore, it allows experiments with various parameters, such as fiber lengths, fiber widths, and fiber density.

The simulated images are generated by superimposing simulated fibers onto a simulated background. The generation of the fibers and the generation of the backgrounds are discussed in Sections 2.1 and 2.2, respectively.

2.1 Background generation

Background in documents is considered to be noise. For authentication purposes, a document is a binary image, where the higher gray level represents fibers and the lower gray level a uniform background. In practice, such a situation does not occur, because backgrounds of real documents are not smooth and they contain a wide gray level range. Based on the background appearance

in real images, the Long-Crested Wave Model [1] is used for background simulation. This model was first introduced in ocean-wave modeling ([7] through [10]).

Suppose that the background is formed of a constant gray level c modulated by a zero mean function $g(x, y)$. Suppose, also, that $g(x, y)$ consists of the sum of a very large number of long-crested waves; i.e.,

$$g(x, y) = \sum_{i=1}^{n} B_i cos(u_i x + v_i y + \phi_i), \tag{2.1}$$

where n approaches infinity, u_i and v_i are x and y components of angular frequency of the ith wave component, and ϕ_i is the phase shift of the same component. The angular frequency of the ith wave component in a direction perpendicular to the crest front is defined by

$$\omega_i = 2\pi f_i = (u_i^2 + v_i^2)^{1/2}. \tag{2.2}$$

The usual assumptions for this model are that

- phases ϕ_i are uniformly distributed over the interval $(0, 2\pi]$ and
- amplitudes B_i are random variables (for detail see [1]).

Since $g(x, y)$ is the sum of a large number of independent random processes, it follows that according to the central limit theorem, $g(x, y)$ is a Gaussian process with zero mean and variance σ_g^2. Variance is calculated as the sum of the variances of all the processes.

Such a model is applicable to ocean-wave modeling but is not practical for image generation because it requires a large number of long-crested waveforms.

To overcome this difficulty, the narrow-band noise model was introduced [1]. A typical narrow-band waveform is given by

$$n(x) = V(x)cos(\omega_c x + \phi(x)), \tag{2.3}$$

where ω_c is the central frequency in the narrow spectrum of the signal $n(x)$, while $V(x)$ and $\phi(x)$ slowly vary the envelope and phase of the signal. A large number of narrow-band noise waveforms form a background, $t(x, y)$:

$$t(x, y) = c + g(x, y) = c + \frac{1}{n}\sum_{i=1}^{n} B_i n(u_i x + v_i y + \phi_i). \tag{2.4}$$

As mentioned previously, $t(x, y)$ is a Gaussian process because it consists of a very large number of independent processes. The narrow-band model can be modified by introducing multiplicative random noise in the following way:

$$t(x, y) = c + g(x, y) = c + \frac{1}{n}\sum_{i=1}^{n} B_i m(u_i x + v_i y + \phi_i), \tag{2.5}$$

where $m(x)$ is defined as

$$m(x) = [(1 - \alpha) + \alpha R]n(x), \qquad 0 < \alpha < 1, \tag{2.6}$$

where R is a random variable uniformly distributed over the range [0,1]. In such a model, n in the above summation need not be a large number and in practice is between 2 and 15.

In the implementation of the long-crested wave model for the background simulation used in this thesis, the following set of parameters is specified:

- image size (maximum 512×512; we deal with images of the size 256×256)

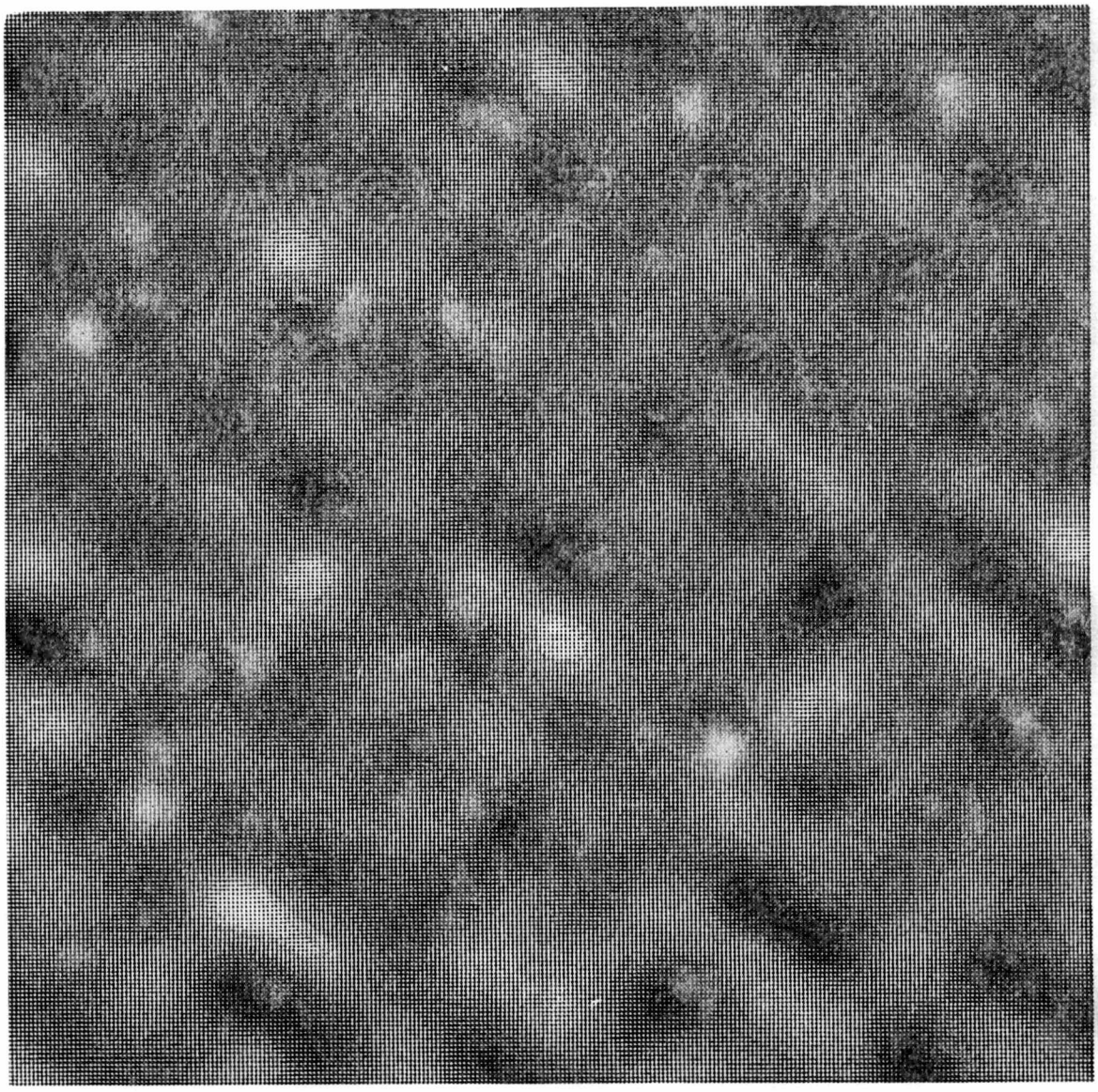
- seed or the initial value of the random number generator (an integer chosen arbitrarily)
- minimum and maximum intensity of the background (because we deal with 256 gray levels, from 0 to 255, minimum and maximum intensity of the background are chosen to be 0 and 200, respectively; background intensity is a normal random variable between specified limits)
- relative frequency, f, of waves (recommended values: from 0 to 1.2)
- number of wave patterns, n, (recommended values: from 2 to 15).

Examples of images with backgrounds simulated by using the long-crested model are shown in Figure 2.1. In figures (a) and (b), the influence of relative frequency on the background image is shown (number of waves, n, is constant); in figures (c) and (d) the influence of the number of vectors on the background is shown (relative frequency is constant).

We may form a very large but still finite number of different background patterns. This number does not approach infinity primarily because of the fact that the random number generator in a computer has a finite period.

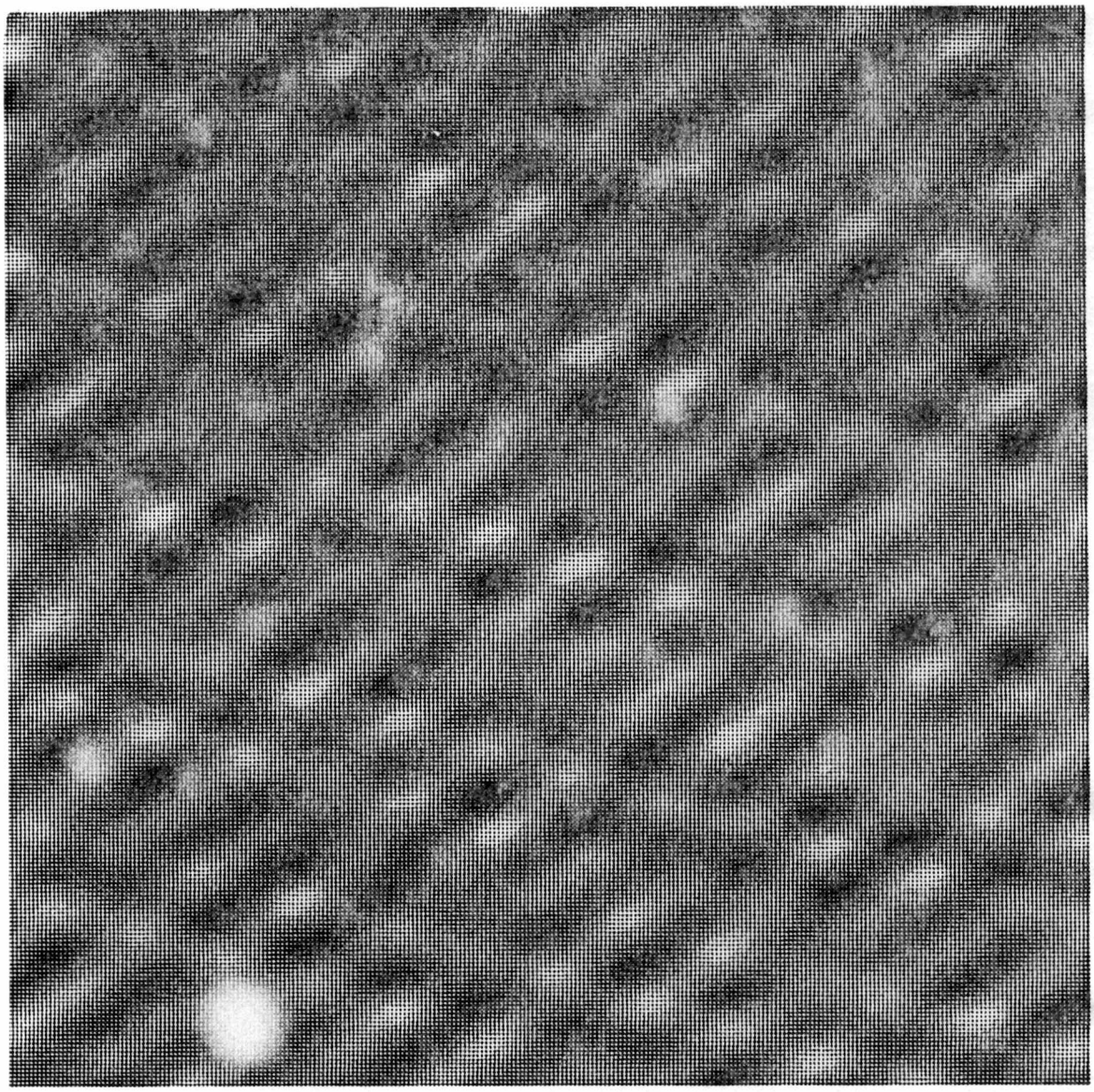
2.2 Fiber generation

In this section, simulation of fibers is discussed. The fibers are modeled by polynomial segments of the first and second order. Specifically, line segments and circle arcs like those shown in Figures 2.2 (a) and (b) are used. All pa-



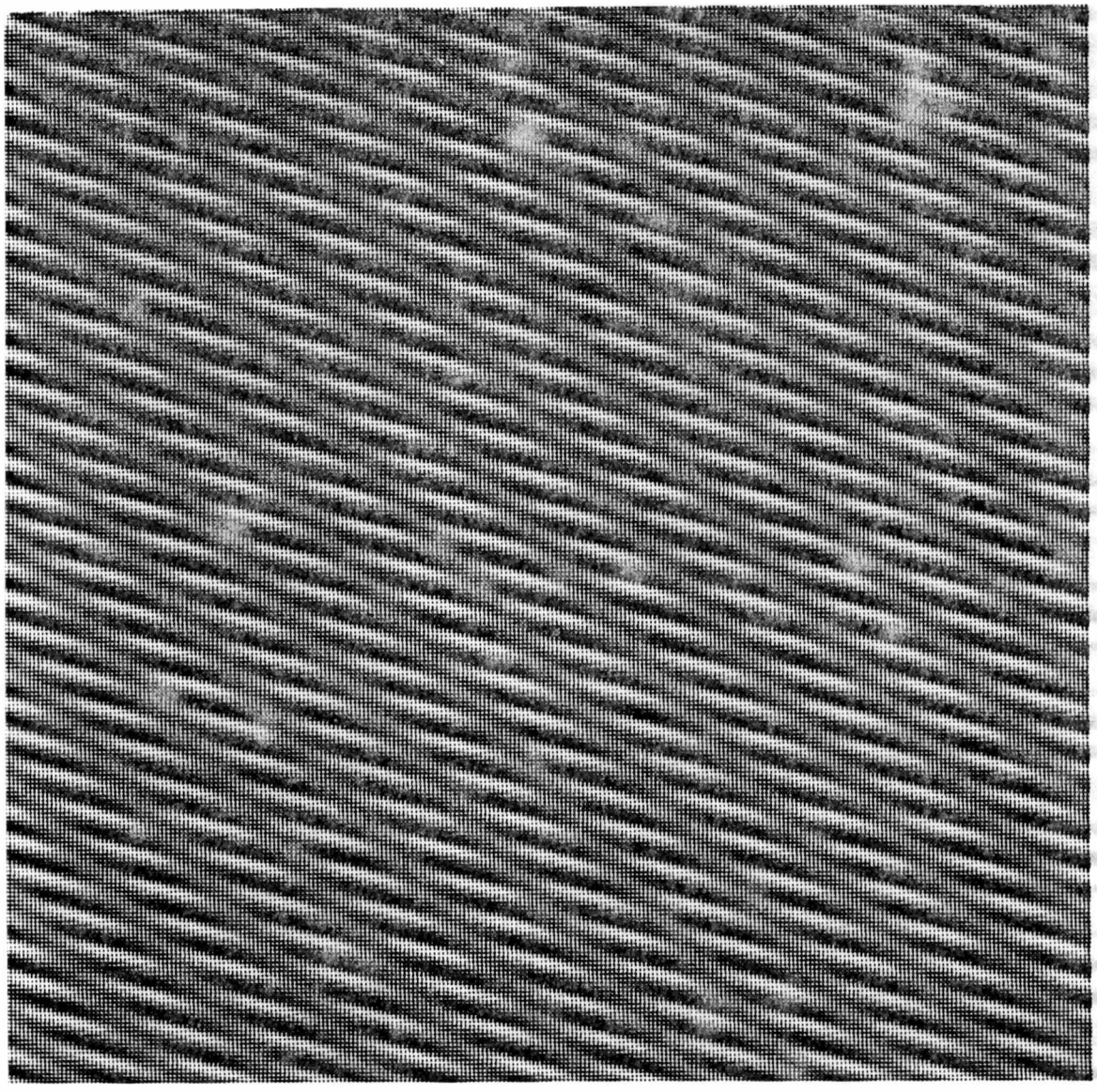
(a)

Figure 2.1: Some examples of backgrounds (image size is 256×256 and minimum and maximum intensities are 0 and 200, respectively): (a) $f = 0.2,\ n = 8$; (b) $f = 0.8,\ n = 8$; (c) $f = 1.0,\ n = 2$; (d) $f = 1.0,\ n = 15$.



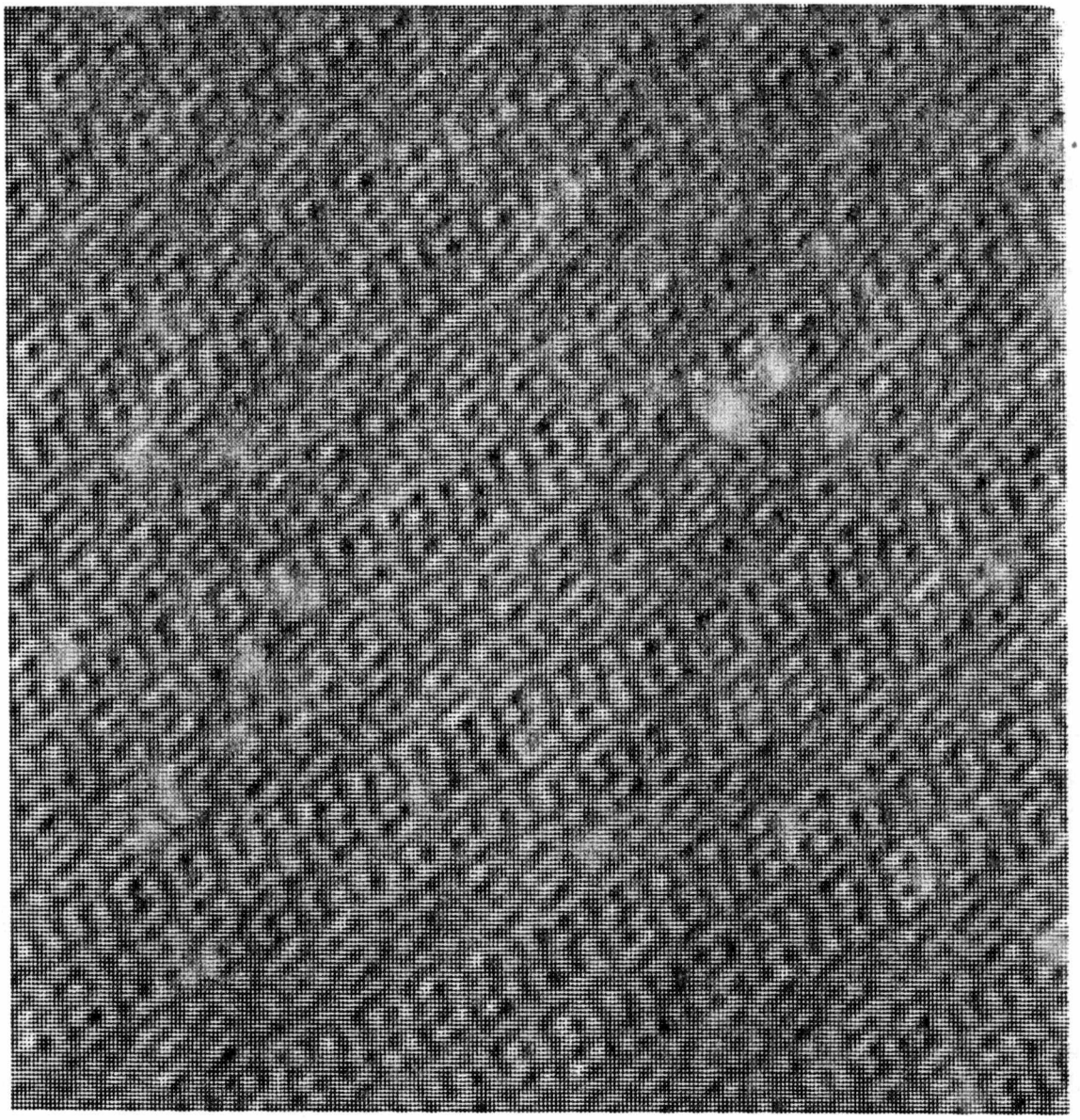
(b)

Figure 2.1: Continued.



(c)

Figure 2.1: Continued.



(d)

Figure 2.1: Continued.

rameters (for circle arcs r, l, and α and for line segments r and α) are random variables normally distributed with changeable mean values and standard deviations. Certainly, fibers can be modeled by higher order polynomials, for example third, fourth, and so on, but this added computational complexity is not needed.

In the present version of the algorithm for fiber image generation, the following set of parameters is specified:

- seed of the random number generator (an integer chosen arbitrarily)
- mean and standard deviation for the fiber intensity that is a normal random variable
- shape of fibers, that is, means and standard deviations for length of fibers, radius of arcs and angle of arcs that are normally distributed random variables (Figures 2.2)
- widths of fibers.

The number of different fiber images that may be formed by a computer is limited by its random number generator and by the image's fiber density.

The following assumptions are introduced when fiber images are generated:

- number of fibers in a document is specified by fiber density per unit area (square inch) and by document dimensions (in inches)
- fibers are randomly distributed over the document and may overlap.

This thesis deals with fibers of one pixel width. For fibers of wider widths, the distribution of gray levels over the fiber cross sections has to be modeled.

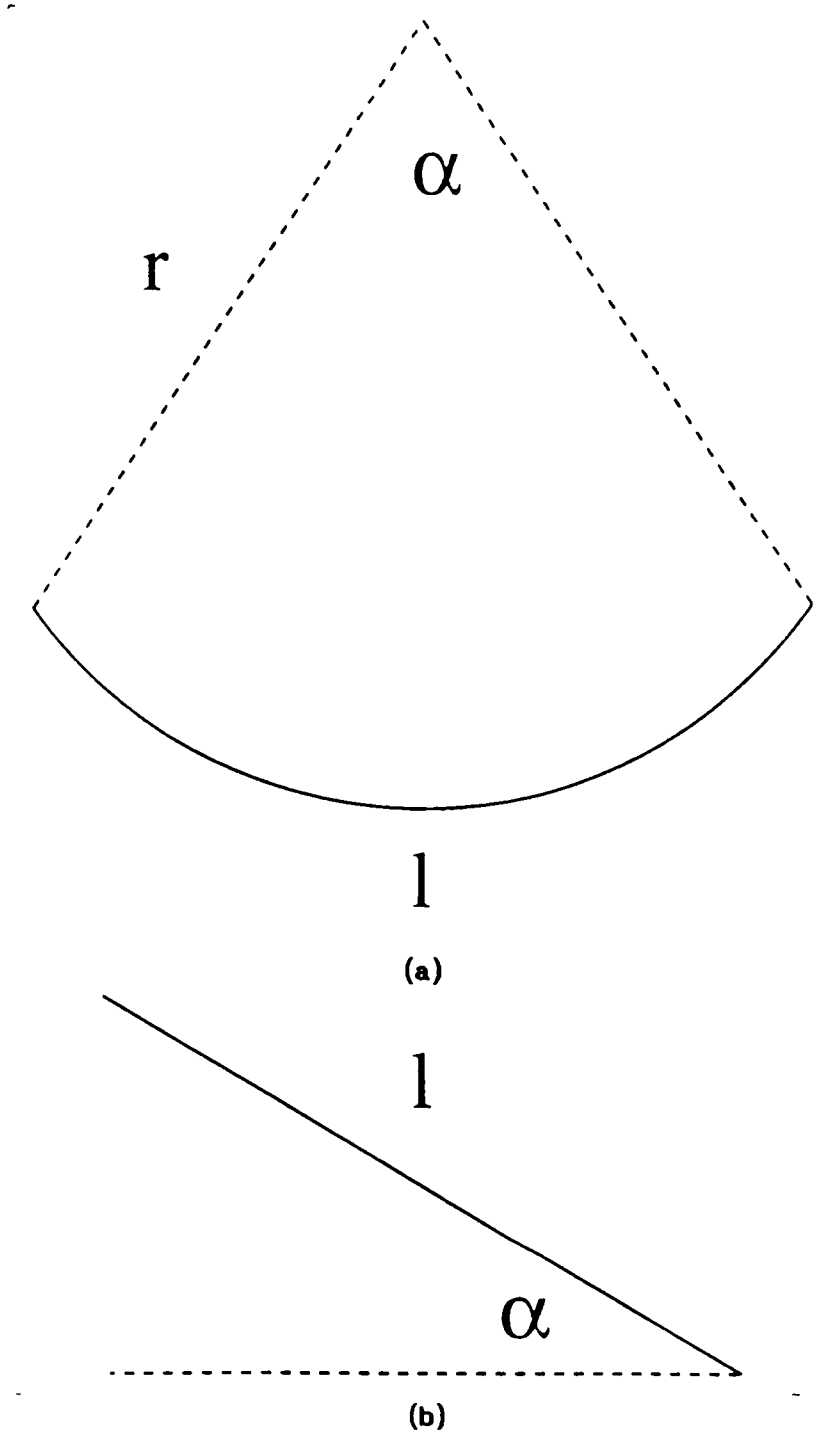
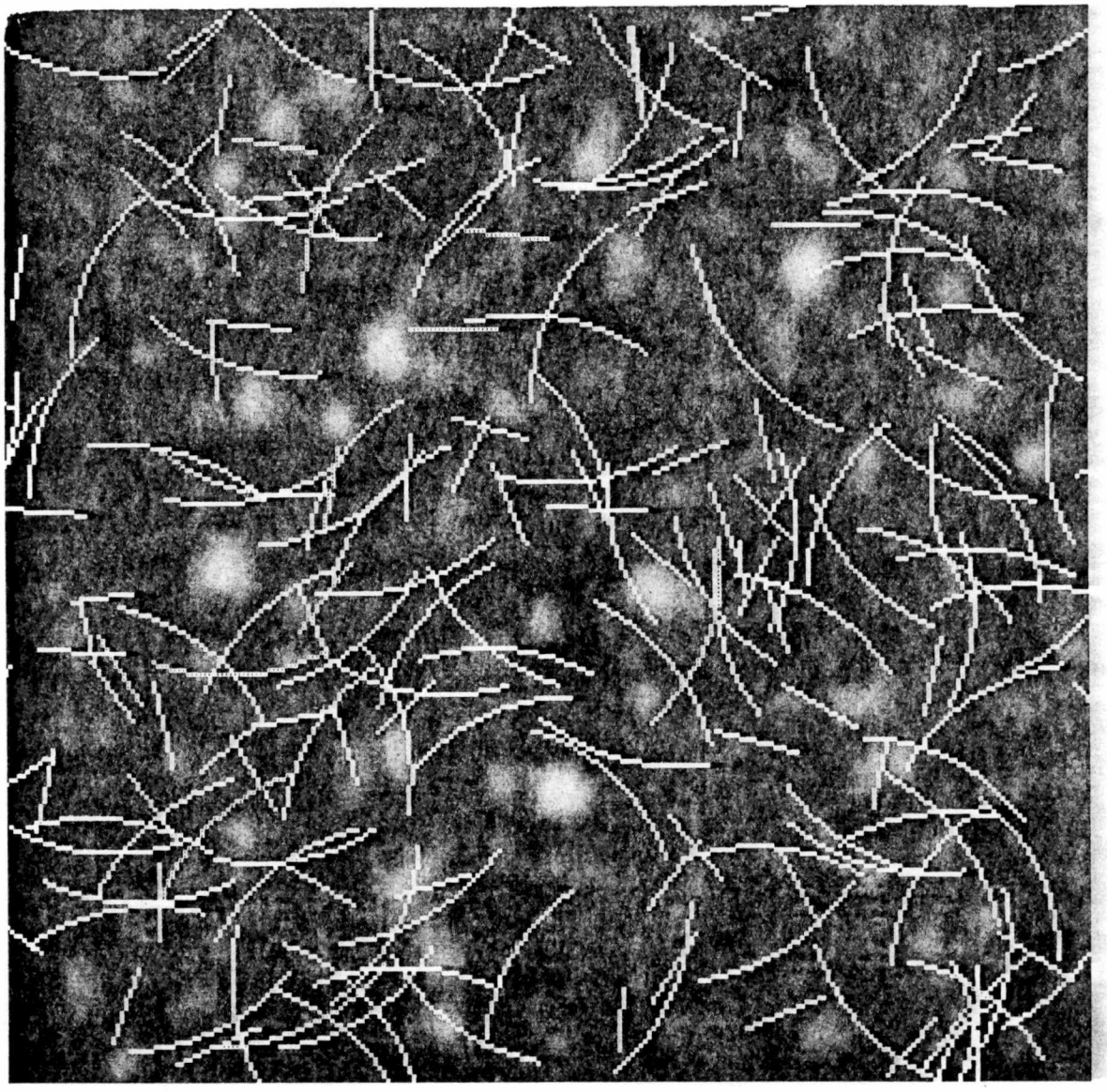


Figure 2.2: (a) A fiber simulated by a circle arc: r, l, and α are random variables normally distributed. (b) A fiber simulated by a line segment: l and α are random variables normally distributed.

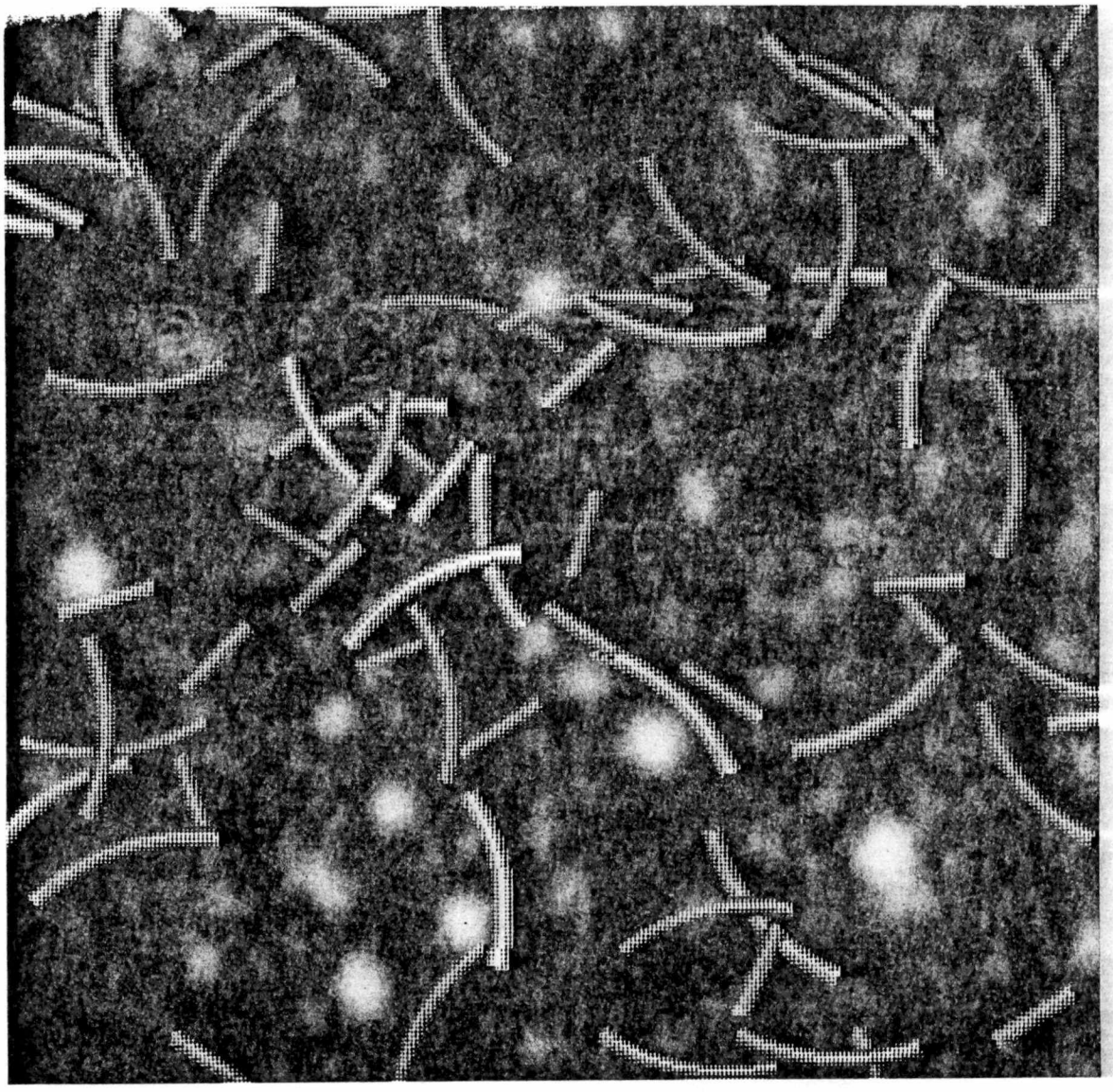
It may be well simulated by using Gaussian functions so that the fiber's central points are the brightest and its peripheral ones the darkest. An example of fiber images with one and five pixel widths is given in Figure 2.3.

Simulated images are generated by superimposing fibers onto backgrounds. Typical examples of synthetic images are shown in Figure 2.4.



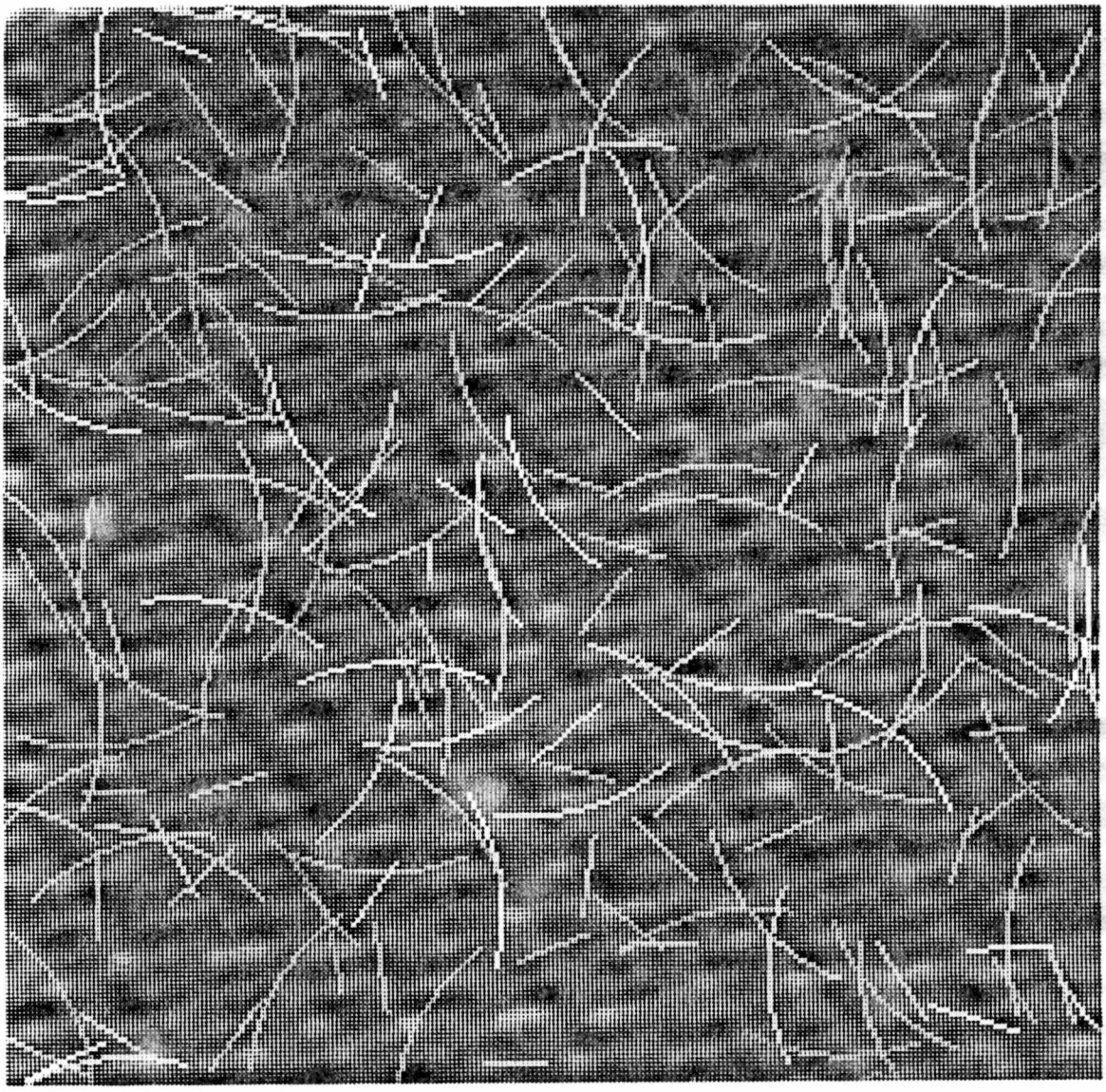
(a)

Figure 2.3: An example of fiber images: (a) one pixel fiber widths and (b) five pixel fiber widths.



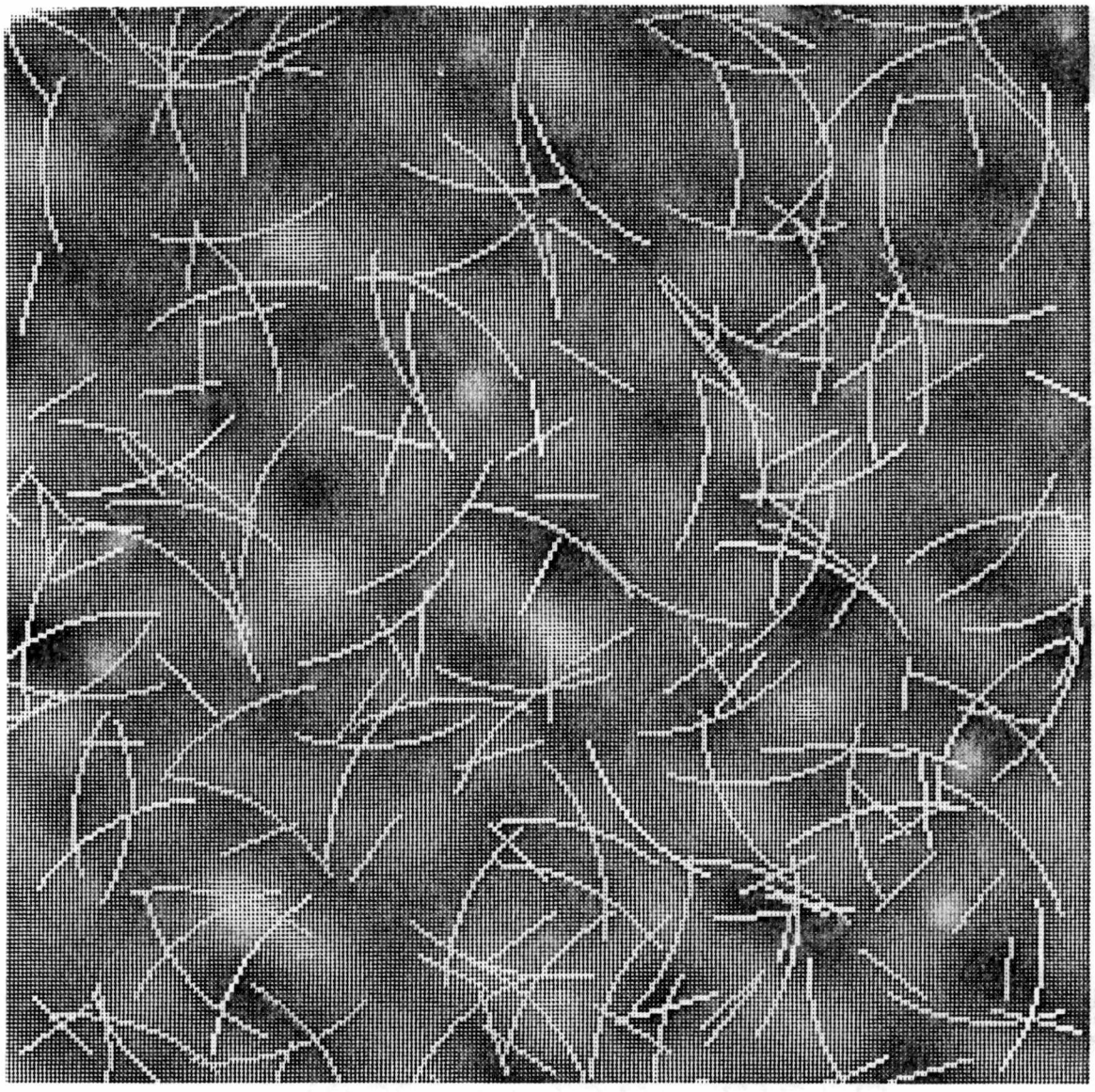
(b)

Figure 2.3: Continued.



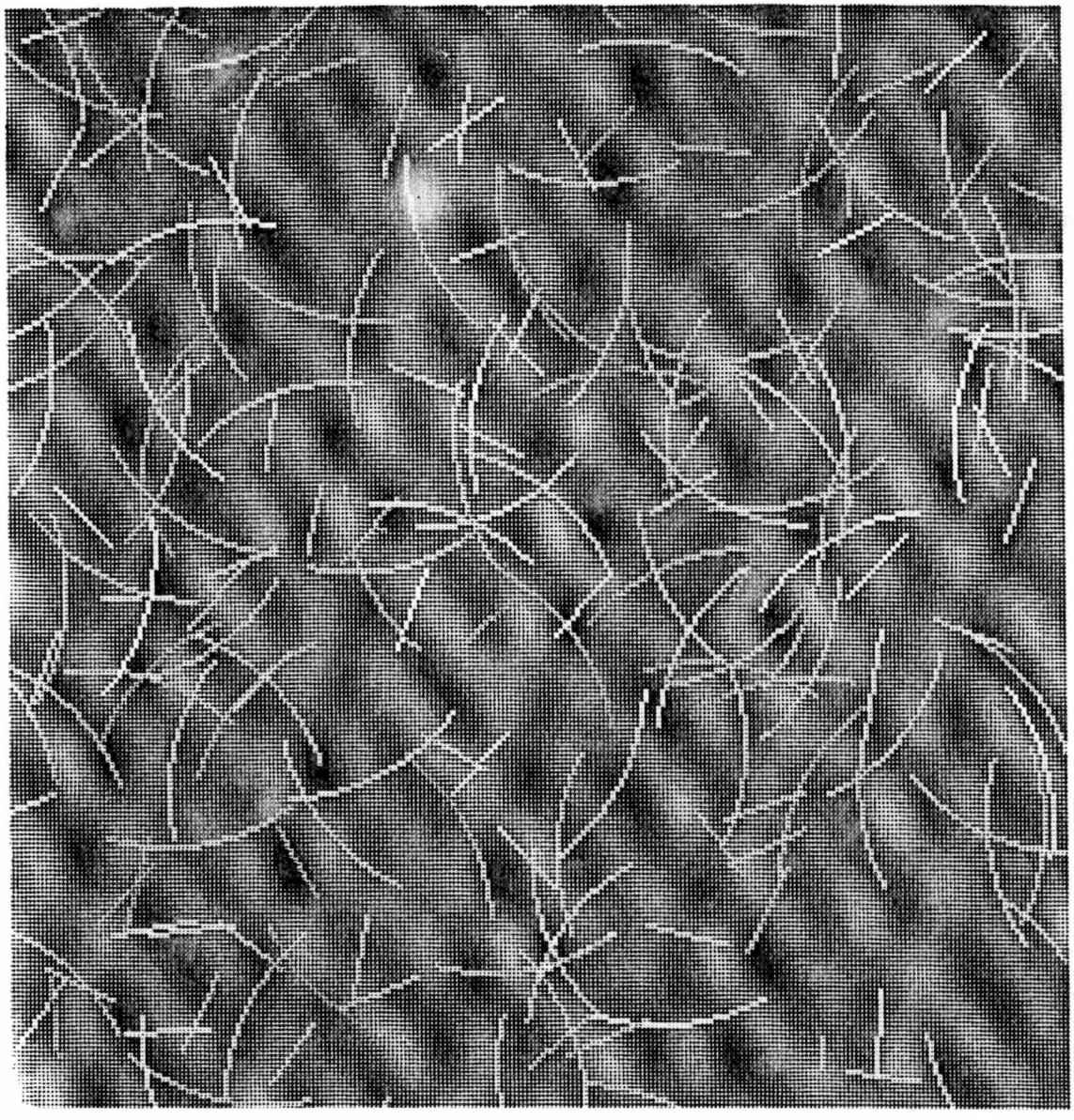
(a)

Figure 2.4: Typical examples of synthetic images.



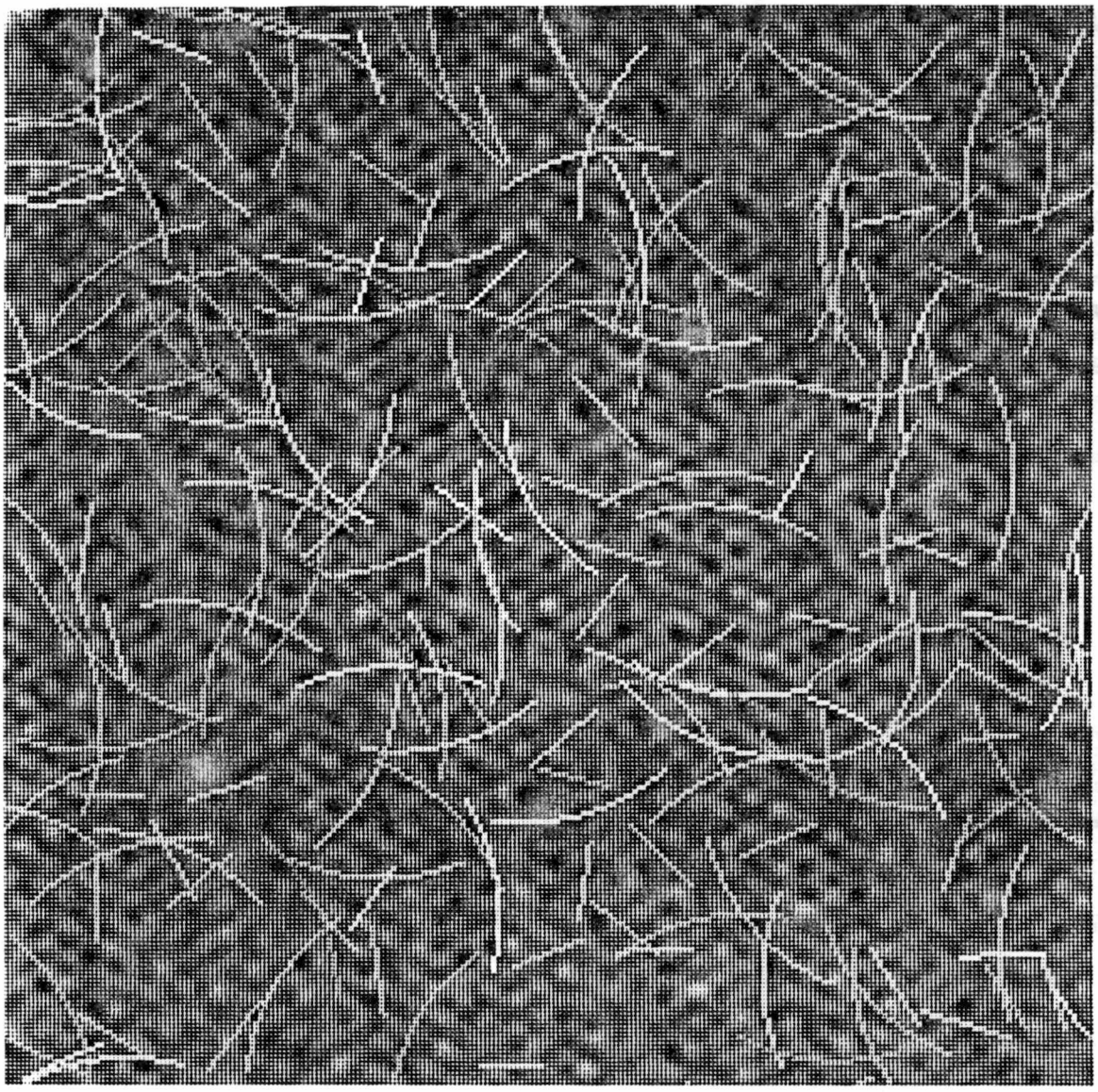
(b)

Figure 2.4: Continued.



(c)

Figure 2.4: Continued.



(d)

Figure 2.4: Continued.

CHAPTER 3

Fiber Extraction

This chapter describes an algorithm for fiber extraction from a document. Also, developed is an algorithm for fiber end points extraction, upon which the authentication procedure is based.

Documents are determined by fibers and their relationships; backgrounds are considered noise. Thus, only fibers in documents are of interest. If all information from documents is to be saved, the following two conditions must be handled by the fiber extraction procedure:

- all fibers from documents are to be extracted properly, and
- backgrounds are to be set to a constant value.

Binary images with bright fibers and dark backgrounds is result from the fiber extraction procedure.

The problem of fiber extraction would be trivial if background gray levels were significantly less than gray levels of fibers. In such a case, an appropriately introduced threshold could solve the problem (above the threshold are fibers; below the threshold is noise). In practice, the situation is different because fiber and background gray levels partly overlap. In Figures 3.1, 3.2 (a), and 3.2 (b) an original document and histograms of the background and fibers of the document

are shown, respectively. Note that the gray levels of the fibers and those of the background are overlapped (compare (a) and (b), Figure 3.2). This means that a simple thresholding technique cannot be used to extract fibers from noisy documents.

In this thesis, fiber extraction is accomplished in two steps:

- step 1: compression of the dynamic range of a background
- step 2: fiber extraction by using an edge detector.

Detection of fiber end points is considered after the fibers have been extracted properly.

3.1 Dynamic range compression

A gradient operator is an edge detector that is very sensitive to intensity changes in images. In documents, abrupt intensity changes almost always occur when fibers are present. However, sometimes there are steep gray level changes in the background, and values of the gradient at these positions are fairly large. Because they may cause incorrect fiber detection, background gray level changes are to be compressed and consequently the background influence on the gradient image of a document.

Suppose that fibers are to be extracted from the document shown in Figure 3.1. When the gradient technique was applied directly on the document, the gradient image, shown in Figure 3.3 resulted. Note that fibers in the resulting

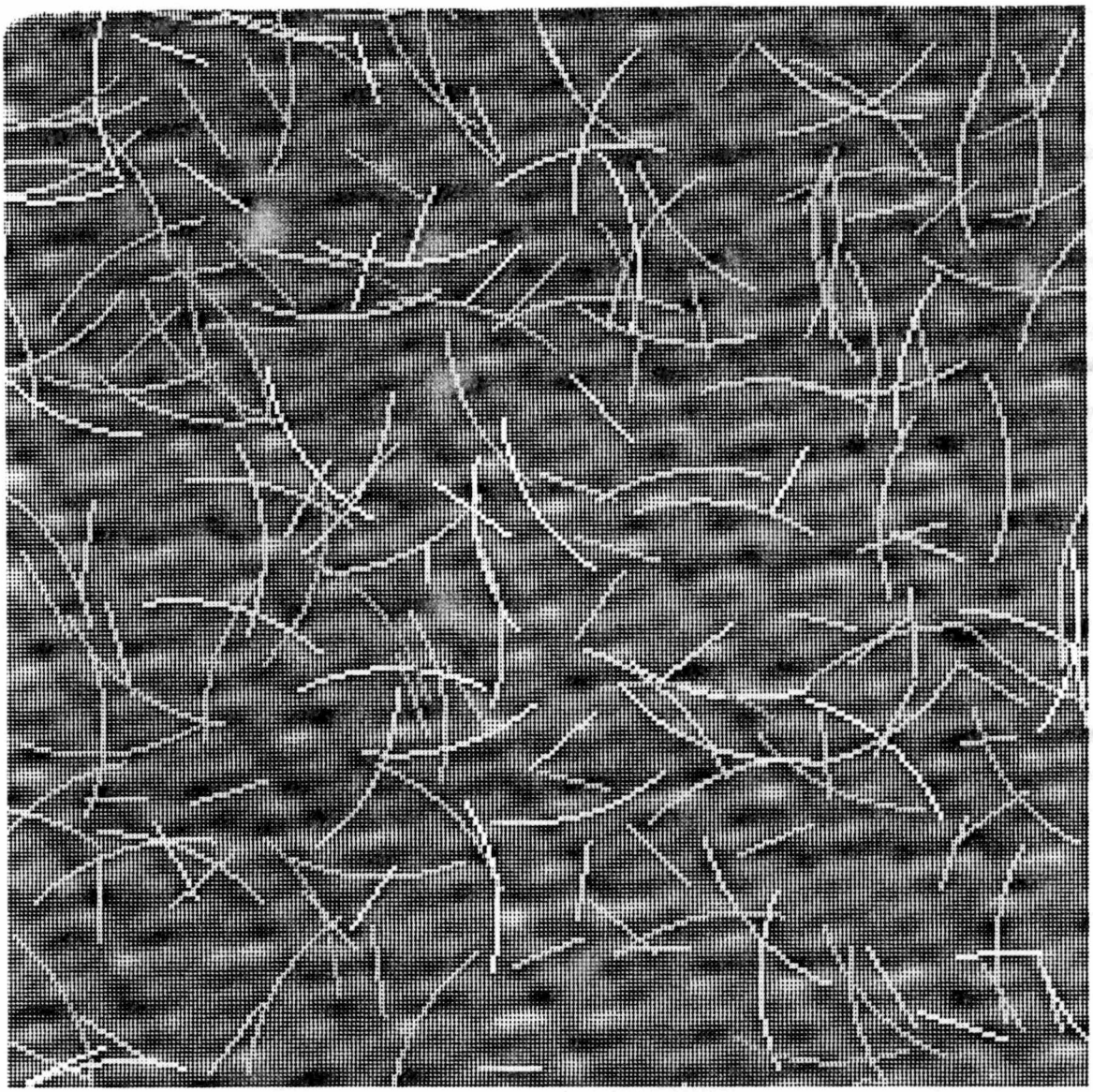
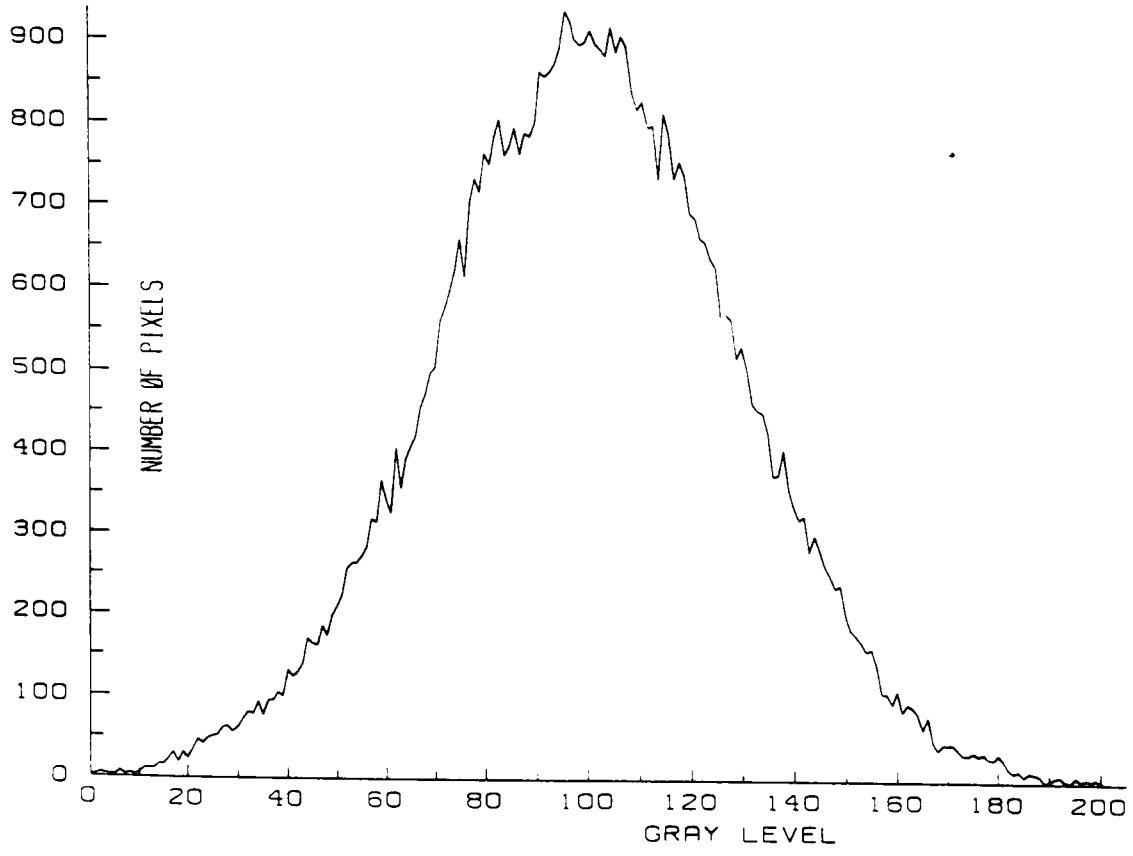
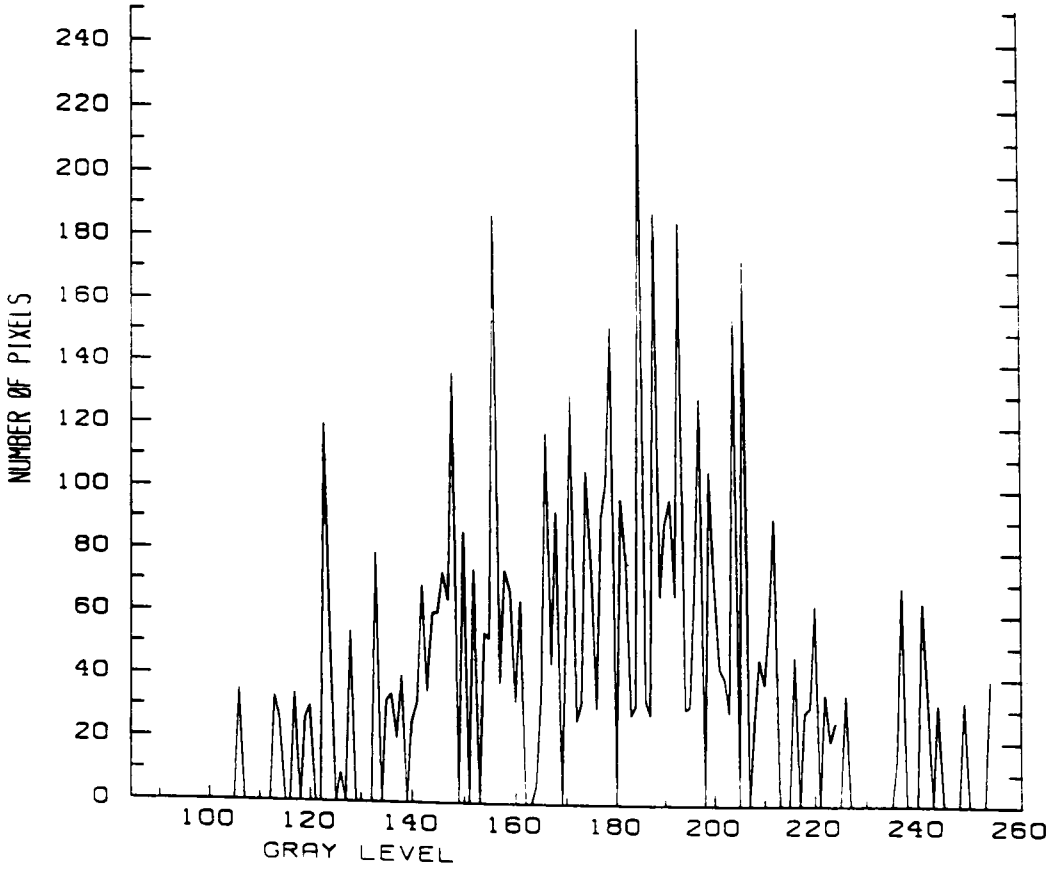


Figure 3.1: An original document.



(a)

Figure 3.2: (a) Histogram of the background of the image in Figure 3.1. (b) Histogram of fibers of the image in Figure 3.1.



(b)

Figure 3.2: Continued.

image are hardly distinguishable from the background; i.e., the image obtained by direct application of the gradient technique is quite useless.

From the histogram image, Figure 3.2, we see that most of the background pixels are of gray levels less then 170. To reduce the background influence on the gradient image, the threshold $T = 170$ is introduced. Pixels with gray levels below the threshold are set to 170 and everything else is left unchanged. The result of this simple transformation is a preprocessed image with compressed dynamic range of the background. The preprocessed image is shown in Figure 3.4.

The gradient image of the preprocessed image is shown in Figure 3.5. Note that fibers in this gradient image are easily distinguished from the background. Compare Figures 3.3 and 3.5.

3.2 Gradient as an edge detector

The gradient technique is used as an edge detector. Several other techniques, such as a technique based on the Laplacian operator [4] or high-pass filtering [4], [11], may also be used, but the gradient technique was shown to be the most convenient here.

The gradient technique is an image sharpening technique and is primarily used when the edges of an image are to be highlighted. For a given image $f(x, y)$

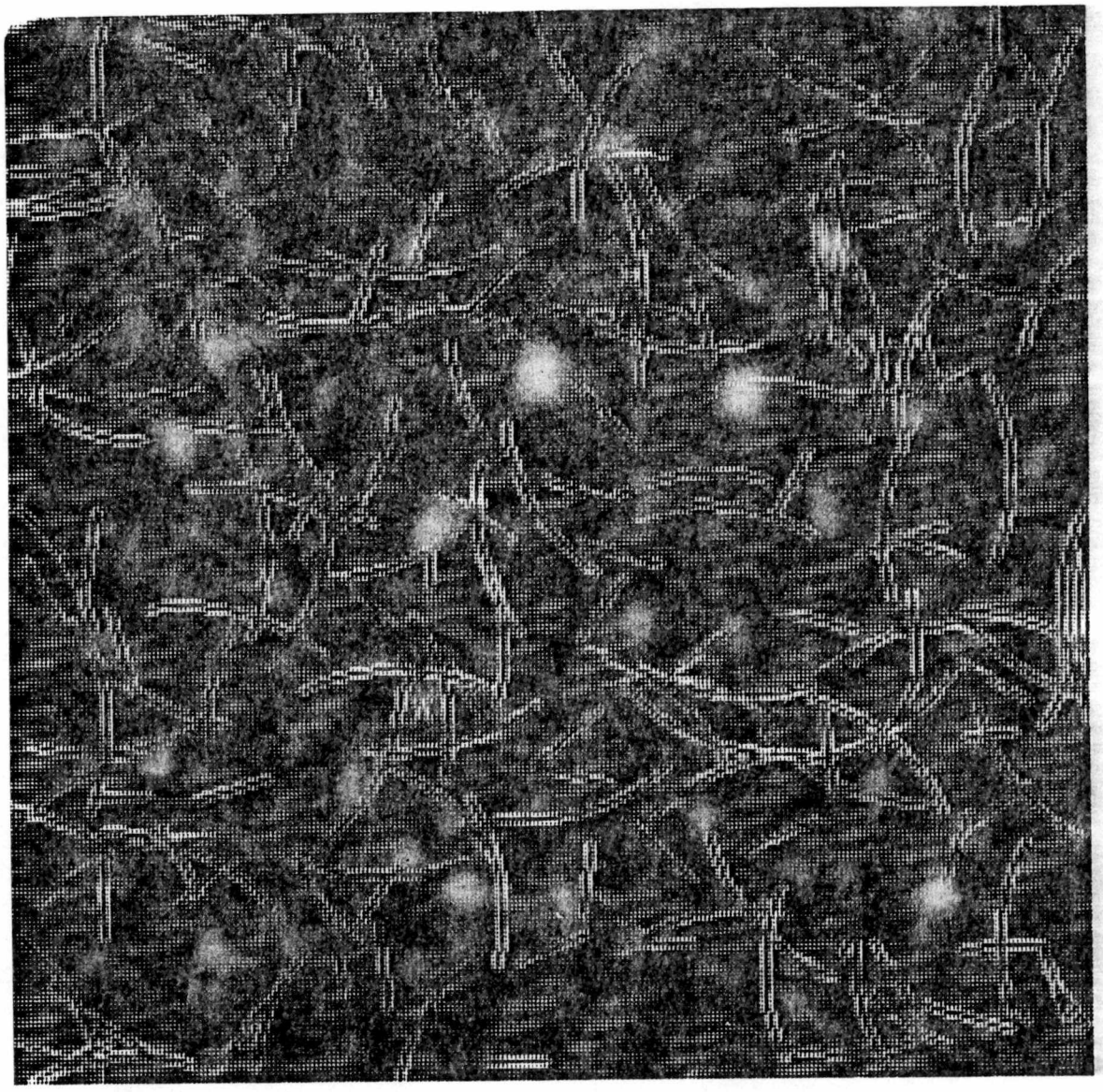


Figure 3.3: Result of the gradient technique applied to the original image in Figure 3.1.

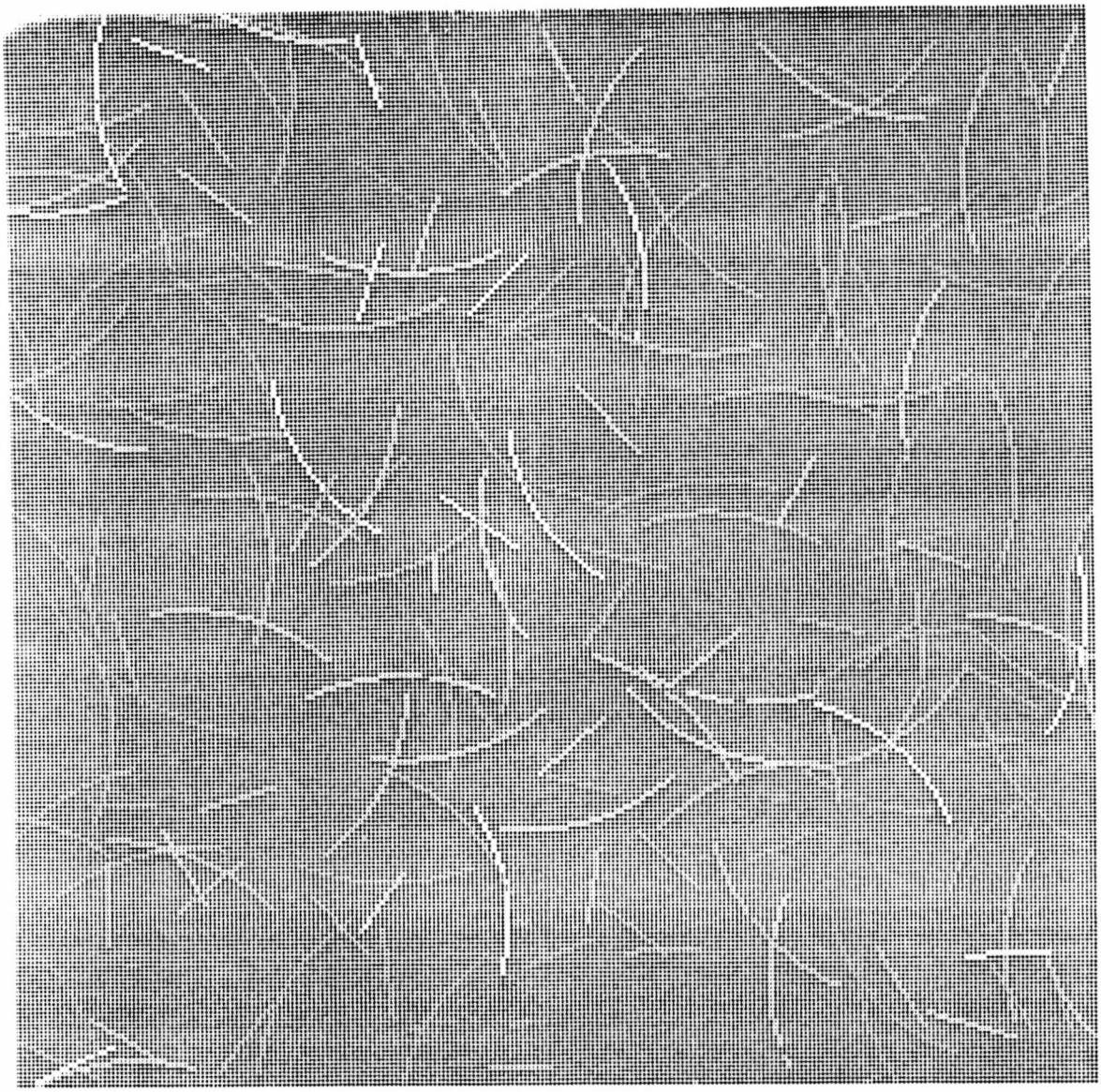


Figure 3.4: Preprocessed image.

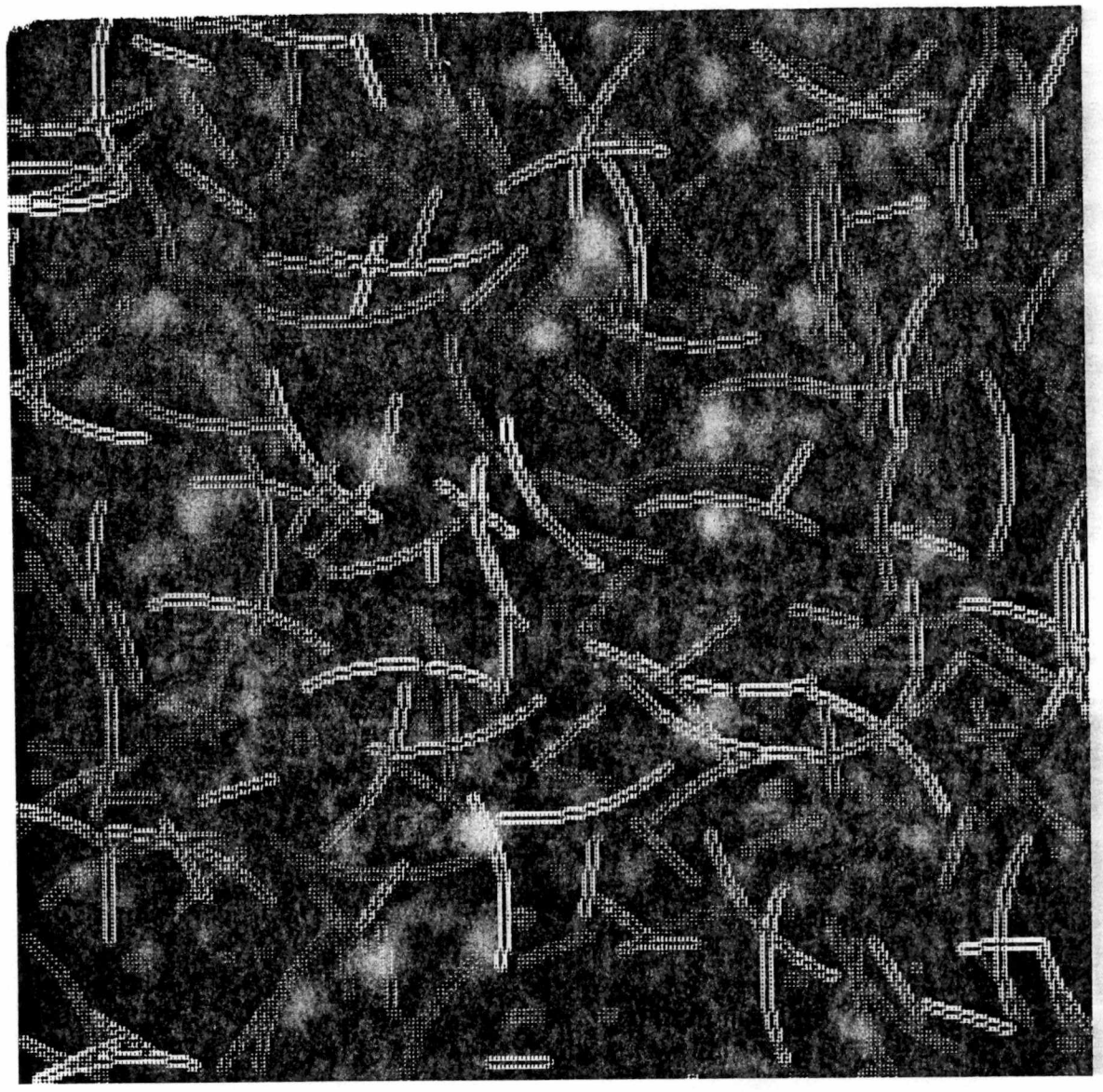


Figure 3.5: Gradient image of the preprocessed image.

the gradient at location (x, y) is a column vector defined as follows:

$$\mathbf{G}[f(x,y)] = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}. \tag{3.1}$$

Gradient vector, $\mathbf{G}$, has the direction of the maximum rate of change $f(x, y)$. In practice, we are usually interested in the magnitude of the gradient, that is, the length of the vector $\mathbf{G}$. The magnitude is defined as follows:

$$G[f(x,y)] = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}, \tag{3.2}$$

or approximately

$$G[f(x,y)] \approx |G_x| + |G_y| = \left|\frac{\partial f}{\partial x}\right| + \left|\frac{\partial f}{\partial y}\right|. \tag{3.3}$$

The magnitude of the gradient (hereafter, simply gradient) is a positive scalar function on which various differentiation techniques are based. This function is a continuous function of two variables. For digital images, it is not possible to calculate the gradient by definition, and the derivatives in Equation (3.2) are approximated by differences. There are several ways to do so, and an approximation of gradient components in the x and y directions, is defined as

$$G_x = [f(x+1, y-1)+2f(x+1, y)+f(x+1, y+1)] - [f(x-1, y-1)+2f(x-1, y)+f(x-1, y+1)] \tag{3.4}$$

and

$$G_y = [f(x-1, y+1)+2f(x, y+1)+f(x+1, y+1)] - [f(x-1, y-1)+2f(x, y-1)+f(x+1, y-1)]. \tag{3.5}$$

Equations (3.4) and (3.5) indicate that the gradient of an image $f(x, y)$ may be computed by using 3×3 masks, as shown in Figure 3.6 (a) and (b). These two masks are known as *Sobel operators*.

An important gradient property is that $G[f(x, y)]$ is zero in regions with constant gray levels, small in regions that are fairly smooth, and large wherever there are abrupt changes in gray levels between adjacent pixels.

The gradient for pixels in the first and last rows and columns of an image cannot be obtained by using Sobel operators. However, if an $n \times m$ gradient for an $n \times m$ image is required. one way to obtain it is to increase the original image to size $(n + 2) \times (m + 2)$. Gray levels of border rows and columns of the enlarged image are determined by gray levels in the adjacent rows and columns in order to avoid abrupt changes in the image to which the gradient method is very sensitive. Then, an $n \times m$ gradient of such an enlarged image may be obtained by using Sobel operators.

It was already mentioned that there are several different gradient approximations. Once the technique is chosen, there are several different ways to create a gradient image. For example, if we want to preserve smooth regions, a threshold may be introduced and all gray levels below the threshold are left unchanged while gradient values above the threshold are set to a certain constant value. Similarly, various other techniques can be introduced [4]. In this thesis, the gradient image is obtained by assigning the magnitude of the gradient to every point (x, y) in the image.

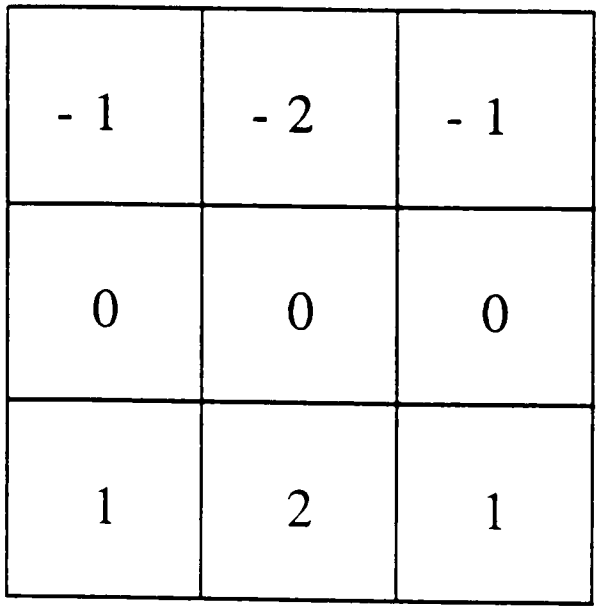


Figure 3.6: (a) Mask used to compute G_x at (x, y). (b) Mask used to compute G_y at (x, y).

3.3 Detection of the end points of fibers

Once fibers are extracted from the noisy background of a document, end points of fibers are to be detected. This is the first step in the document authentication algorithm.

The document authentication procedure is performed between a set of images prestored in a computer (database) and an image under consideration (test image). The test image may be translated and/or rotated relative to the images in the database. Also, some fibers may be missing in the test image. Because some relationships among fiber end points are invariant to image translation and rotation (for example, distances and angles between points), the document authentication procedure is based on fiber end points.

In short, the algorithm for fiber end point detection works as follows. An image is scanned, and when a non-zero pixel is found, it is to be determined whether that particular pixel belongs to a fiber or not. If the pixel belongs to a fiber, it is to be determined whether it is an end point of a fiber or not. All but end points of fibers are *erased* from the document so that the resulting image consists of fiber end points only.

Images are scanned from left to right and from top to bottom by examining every pixel. Suppose that a non-zero pixel is found at location (i, j). Then, the following situations are possible:

- the pixel is a fiber end point
- the pixel is a fiber point but not an end point

- the pixel is a cross section of at least two fibers
- the pixel is noise.

Let us define a 3×3 neighborhood of the pixel at (i, j) (a mask centered at (i, j)). If the pixel is an end point, either one more pixel in the mask is non-zero, or, at most, two more pixels in the mask, in the same direction, are non-zero. Here, *in the same direction* means in the same half of the 3×3 mask. If either one of these cases is satisfied, the search continues in the 3×3 neighborhood of the next non-zero pixel (next to the pixel at location (i, j)) checking the continuation of a fiber. If continuation occurs in three pixels, the point at location (i, j) is considered an end point of a fiber. In all other cases, a pixel at location (i, j) is not considered a fiber end point. These *other* cases indicate that
(1) there is more than one non-zero pixel in the 3×3 neighborhood of (i, j) in $different$ halves of the mask (the pixel at (i, j) is either part of a fiber, noise, or an intersecting point of two or more fibers);
(2) there is just one point in the neighborhood, but there is not a continuation of a fiber in at least three points (the pixel at (i, j) is noise);
(3) there are two points in the neighborhood of the point in the same direction, but the fiber is not continued (the pixel is noise).

The algorithm for end point detection does not detect the end point(s) of fibers that touch other fibers by their end point(s). In Figure 3.7, the end points of fibers of an original document are shown. To gain some insight into

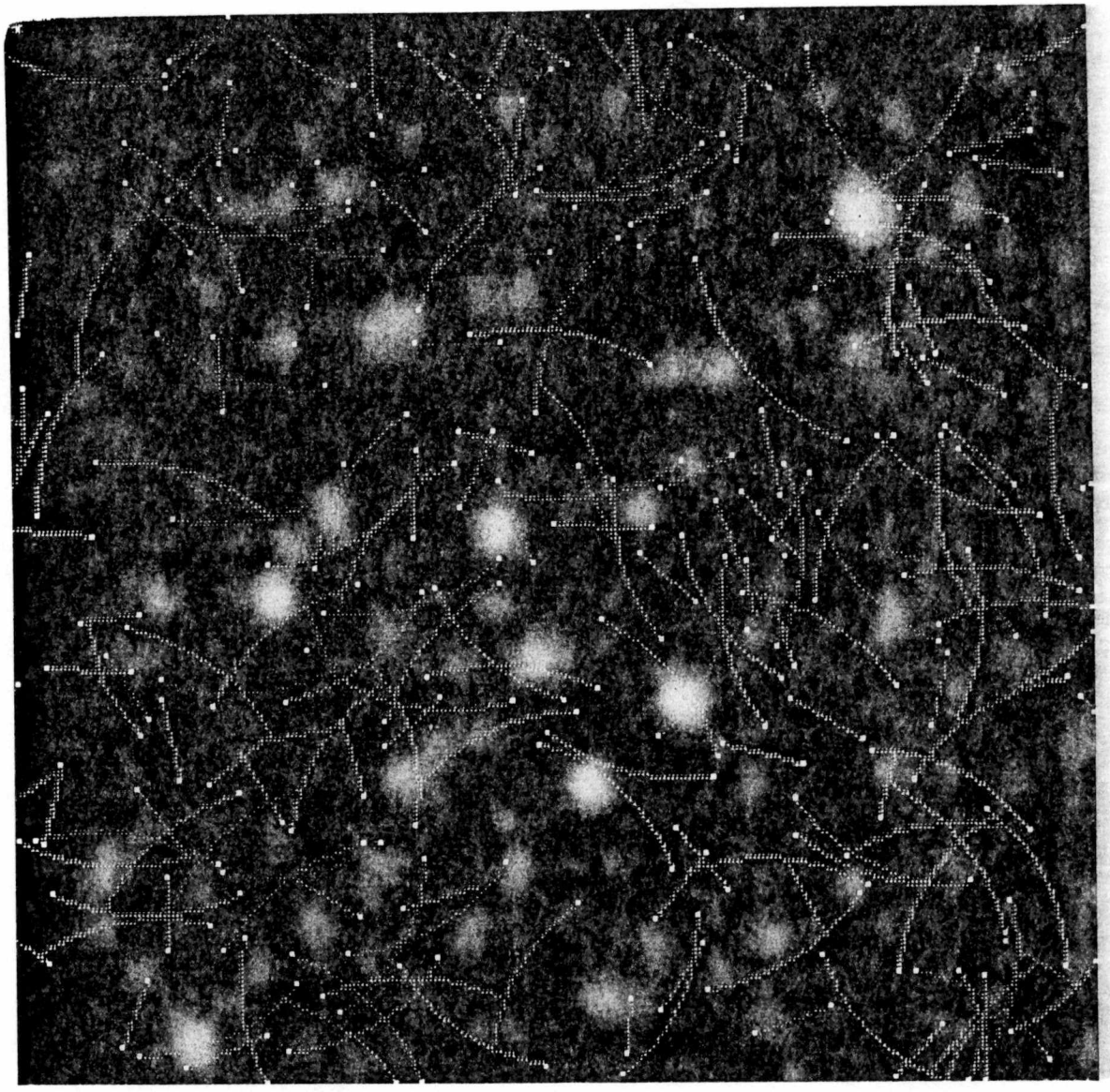


Figure 3.7: End points of fibers of a document superimposed on the document itself. End points were found by the algorithm for end point detection.

the manner in which the program for end point detection works, end points are superimposed on the document itself. Note fibers whose end points were not found. In spite of the fact that *touching* situations are not of crucial importance for the authentication process, one has to keep in mind that the algorithm has no solution for these situations.

CHAPTER 4

Document Authentication

This chapter introduces the concept of distances between fiber end points (hereafter, simply distances) on which document authentication is based. We deal with documents with zero backgrounds on which fibers of one pixel widths are randomly distributed.

The document authentication problem is defined as follows. Suppose that there is a database consisting of a large number of documents and a test image under current investigation. We want to determine whether the test image belongs to the database or not.

The authentication procedure is derived from the following assumptions:

- the test image may be translated relative to database images
- the test image may be rotated relative to database images
- the test image may contain only a subset of fibers relative to database images.

Document authentication is divided into three steps. The first two steps are related to end point images while the third step is related to fiber images:

- step 1: establishing a three point match, i.e., establishing if any three points in the test image match three points of any database image

- step 2: establishing if remaining points in the test image have match in the image chosen in step 1

- step 3: establishing if fiber images of two documents (the test document and one chosen from the database) match.

Translation and rotations are caused by misalignment during image acquisition. Image acquisition is a repeatable process, and during its performance the test image may be translated and/or rotated. Maximum values for translation and rotation are prespecified and denoted by Δx_{max}, Δy_{max}, and θ_{max}. They may be both positive and negative.

There is another effect that has to be taken into consideration in the authentication procedure: the test image is subject to digitization effects; i.e., same points may be detected at different locations when an identical experiment is repeated. Here, *different locations* indicates one of the eight immediate neighbors of considered points.

A test image may contain fewer fibers than the original if such an image is stored in the database. The maximum number of missing fibers, N_{max}, is prespecified. If a whole region of fibers is removed, up to one quarter of all fibers may be missing. Hence, the test image may contain only a subset of original fibers, and, consequently, the end point image of the test image may contain only a subset of original end points.

In this chapter the authentication algorithm is derived. First, the organization of the database is described; than, the matching procedure is described.

4.1 Database generation

There are two types of databases: (1) a database that contains end point images and (2) a file that contains information about specific end points of each image. In the previous chapter, formation of the first database was described. Here, the formation of the second database file is described.

4.1.1 File generation

Suppose that a fiber image has already been created. Then, all images are divided into four quadrants (see Figure 4.1), and for every quadrant, J points are chosen as follows:

- the points closest to the physical center of a quadrant
- each of J points belongs to a different fiber.

J points are chosen as described because
(1) center points remain in images after rotation and translation (points close to the physical edges of images may disappear) and
(2) if one of the center fibers is missing, only one of the J points is missing.
The number J is determined by the maximum number of missing fibers in an image

$$J = N_{max} + 1, \tag{4.1}$$

where N_{max} denotes the maximum number of missing fibers.

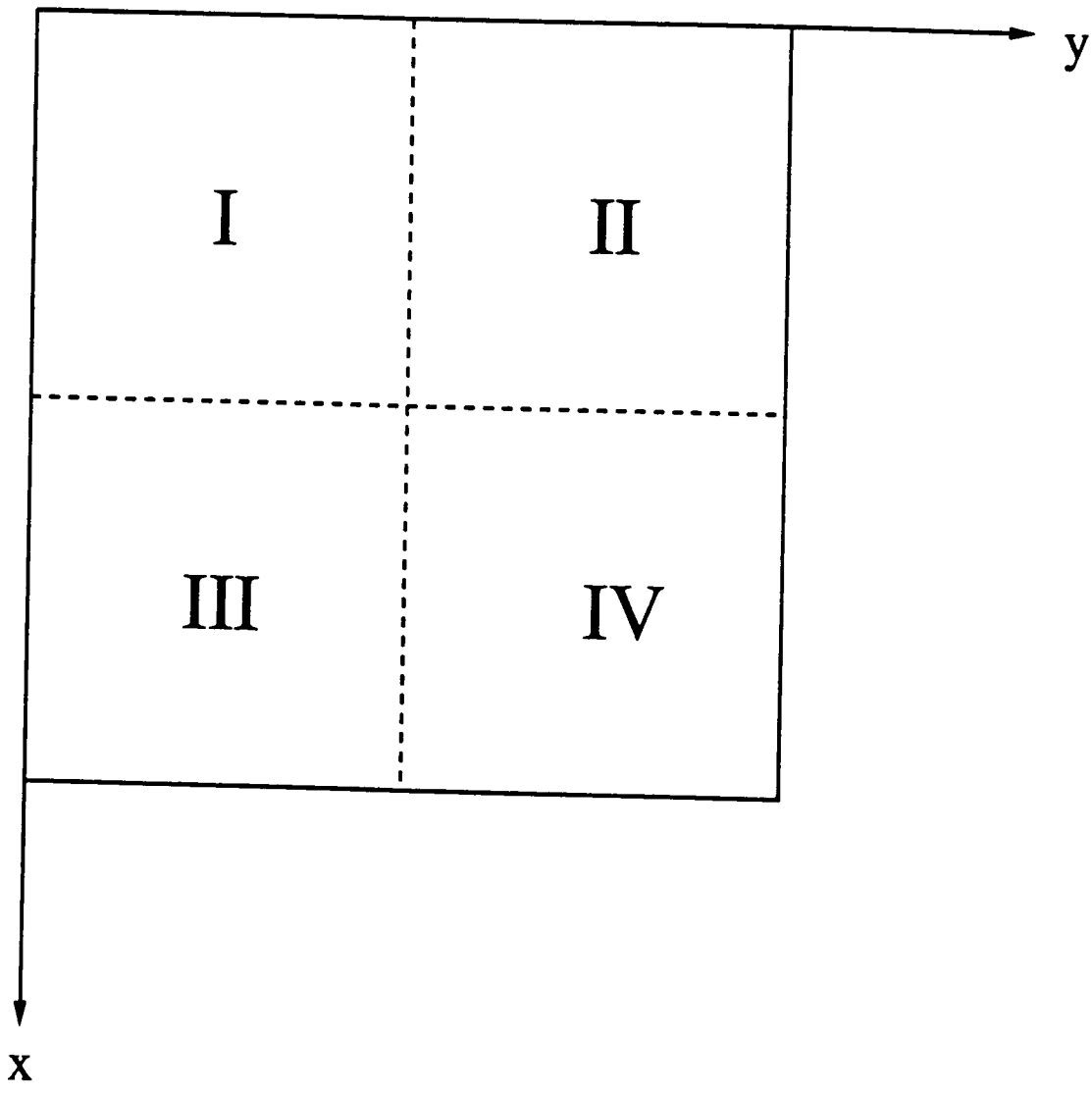


Figure 4.1: An image divided into four quadrants.

Information pertaining to a particular image is organized in the database as follows:

- Information pertaining to the J points of the first quadrant is followed by that of the second, third, and fourth quadrant, respectively.
- For each of the J points of any quadrant, the following information is stored:
  - the physical coordinates of a point, x and y
  - a sequence of distances between the considered point at location (x, y) and all points within the circle of a prespecified radius R centered at (x, y), organized in an ascending order.

4.2 Matching

In this section we establish a procedure that determines the presence of a match between the test image and an image in the database. All images that are considered are end point images.

If two images under consideration are matched, every point from the test image must have a corresponding point in the database image. This statement coincides with the previously introduced assumption that points in the test image may be only a subset of database image points. This condition is used in the first, preliminary examination, if a database image is a potential match. If the number of points in the test image is greater than the number of points

in the considered database image the matching procedure stops and the next database image is considered; otherwise, the matching procedure continues for the chosen database image. If the matching procedure continues, it is divided into two steps:

(1) establishing if any three points in the test image match three points of any database image, and

(2) establishing if the remaining points in the test image match the image chosen in (1).

Step (1) determines rotation and/or translation between the test image and the database image. Step (2) determines a correspondance between end point images of two documents. If both steps, (1) and (2), are successfully done, the database image is labelled as a possible match. In Section 4.4, the criterion for assigning a match to the labelled database image is described. If the test image and labelled database image match, the authentication procedure is successfully completed. Otherwise, it continues until all images in the database are examined or until a match is established.

4.3 Point matching

Point matching establishes if any three points in the test image match three points in an image in the database.

The point matching procedure is divided into three steps, one step for each pair of points. If the first point match is found, the procedure continues for

the second point match. Analogly, if the second match is found, the procedure continues for the third point match. If any of the three steps fails, the procedure continues by attempting the first match between the test image and *all* images in the database for all points not considered before the last match, first, second or third, has been established.

The objective of the point matching procedure is to determine the rotation angle θ and the translation parameters, x_0 and y_0, between the test image and the database image. Theoretically, to achieve this goal we need two instead of three pairs of matching points, required by the point matching procedure. Indeed, if x and y denote coordinates in the old coordinate system and x' and y' denote coordinates in the new coordinate system, the transformation formulas between two coordinate systems involving translation and rotation, are

$$x' = (x - x_0)cos\theta \ + \ (y - y_0)sin\theta \tag{4.2}$$

and

$$y' = (y - y_0)cos\theta \ - \ (x - x_0)sin\theta. \tag{4.3}$$

From Equations (4.2) and (4.3) it is clear that three unknowns (the rotation angle and the translation vector, determined by translations in x and y directions) may be determined from two pairs of matching points (four equations, three unknowns). However, this is true for the continuous case, only. For the discrete case, the accuracy obtained by using Equations (4.2) and (4.3) is limited because of the finite pixel size. Unfortunately, this accuracy does not satisfy the document authentication requirements.

To overcome the accuracy problem, the three point matching procedure is established. Furthermore, to increase accuracy of computation, three points far apart in an image, i.e., in different quadrants are chosen. This is recommended since the authentication algorithm deals with pixel accuracy, e.g., integer numbers.

In the next three subsections the point matching procedure for every pair of pixels is described. In short, each of the three pairs of matches is obtained in a different way. The first match is considered in the first or second quadrant, since one quadrant may be missing. This match is considered in the area that is determined by prespecified maximum translation and rotation parameters. The second match is considered in the second or third quadrant but only on the circle which is determined by the first match. Finally, the third match is considered in the third or fourth quadrant in the 3×3 neighborhood of the point where the third match point is expected. Thus, establishing the first match has the largest degree of freedom. The second match has a smaller degree of freedom, while the third match is considered in the 3×3 neighborhood of an expected point, merely because of the digitization effect.

4.3.1 Determining the first point match

The first point match is established by using prespecified values for maximum rotation and translation. Suppose that there is a point P at location (x, y) in the test image. In a database image, corresponding point P' at (x', y') may be

displaced, at most, by

$$|\Delta x| = |x - x'| = |(cos\theta_{max} - 1)x' - y'sin\theta_{max} + \Delta x_{max}| \tag{4.4}$$

and

$$|\Delta y| = |y - y'| = |(cos\theta_{max} - 1)y' - x'sin\theta_{max} + \Delta y_{max}|, \tag{4.5}$$

where θ_{max}, Δx_{max}, and Δy_{max} denote maximum translation and rotation parameters. Note that maximum displacements $|\Delta x|_{max}$ and $|\Delta y|_{max}$ are obtained for upper left and lower right corners of quadrants. They are functions of the image size.

Relative to Equations (4.4) and (4.5), a point $P_1'(x_1', y_1')$ in a database image may be considered a possible match for point $P(x, y)$ in the test image if the following two conditions are satisfied:

$$|x - x_1'| \leq |\Delta x| \tag{4.6}$$

and

$$|y - y_1'| \leq |\Delta y|. \tag{4.7}$$

Point P' is a match for P if, in addition to Equations (4.6) and (4.7), the sequence of distances for P correspond to the sequence of distances for P'. Correspondence here means:

- two sequences may differ from one another in a prespecified number of distances (determined by the maximum number of missing fibers) and
- a distance D_1 from the first sequence corresponds to a distance D_2 from the second sequence if $|D_1 - D_2| \leq T_d$, where T_d is the threshold value de-

termined by the digitization effect; theoretically, T_d is zero, but in practice it is plus or minus one.

Recall that a set of distances for a point P is formed as an ascending order of distances from P to all points within a circle of the prespecified radius, R, centered at P.

Two sequences are compared for all of the J points in corresponding quadrants for which conditions (4.4) and (4.5) are satisfied. An image by image comparison is made throughout the whole database. When the first point match is found for an image in the database, the second point match is attempted for the same image. Note that the first point match is tried for the first or for the second quadrant only. The second quadrant is considered when all J points in the first quadrant are missing.

4.3.2 Determining the second point match

Suppose that the first point match was found for the test image point (x_1, y_1) and for the database image point (x'_1, y'_1) (let us denote that image as the $K-th$ image in the database). The two points are from the same quadrants in images, say $m-th$ quadrant, where m is 1 or 2.

Once the first point match is found, the second match is attempted in the $(m+1)-st$ quadrant of the $K-th$ database image (i.e., second or third quadrant). Suppose that there is a point at location (x_2, y_2) in the test image in the $(m+1)-st$ quadrant. The corresponding point (x'_2, y'_2) in the database

image, if any, should lie on the circle

$$(x - x_1)^2 + (y - y_1)^2 = D^2, \tag{4.8}$$

where D is determined by

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{4.9}$$

This is true theoretically, but in practice, because of the digitization effects, the second point match is to be attempted on an annulus rather than on the circle. The annulus is defined as

$$(D - |T_d|)^2 \leq (x - x_1)^2 + (y - y_1)^2 \leq (D + |T_d|)'^2, \tag{4.10}$$

where T_d, as earlier, denotes the digitization threshold value. Theoretically, T_d is zero, but in practice it is plus or minus one. In other words,

$$\sqrt{(x_1' - x_2')^2 + (y_1' - y_2')^2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} + T_d, \tag{4.11}$$

where the digitization effects are compensated by the threshold T_d.

If there is a point on the annulus, it still does not mean that it is a match. If the point is a matching point, the set of distances for (x_2', y_2') must correspond to the set of distances for (x_2, y_2). When both of the conditions are successfully met, the second point match is established.

4.3.3 Determining the third point match

When both first and second matches are established for the $K - th$ database image, the third match is attempted for the same image. If the first match is

found in the $m-th$ quadrant, the second one in the $(m+1)-st$ quadrant, the third match is attempted in the $(m+2)-nd$ quadrant. Equivalently, if m is 1 or 2, the third match is attempted in the third or fourth quadrant.

Suppose that there is a point at location (x_3, y_3) in the $(m+2)-nd$ quadrant of the test image. A database point (x'_3, y'_3) is a third match if the following are satisfied:

- $|x_3 - x'_3| \leq |\Delta x|$ and $|y_3 - y'_3| \leq |\Delta y|$
- $$\sqrt{(x'_1 - x'_2)^2 + (y'_1 - y'_2)^2} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} + T_d \qquad (4.12)$$
$$\sqrt{(x'_1 - x'_3)^2 + (y'_1 - y'_3)^2} = \sqrt{(x_1 - x_3)^2 + (y_1 - y_3)^2} + T_d \qquad (4.13)$$
$$\sqrt{(x'_3 - x'_2)^2 + (y'_3 - y'_2)^2} = \sqrt{(x_3 - x_2)^2 + (y_3 - y_2)^2} + T_d, \qquad (4.14)$$
  where T_d is plus or minus one
- distance sequences for two points correspond to each other.

Before leaving this section let us make one more remark. Suppose that the first match is established for the $K-th$ image in the database. Then the second match is attempted for the same image. If the second match cannot be established, the point matching procedure continues by attempting the first match between the test image and all images in the database (including the image K) for points not considered before the first match has been established. If the second match for the $K-th$ image in the database is established, the third match for the same image is attempted. If the third match fails, the

point matching procedure continues by attempting the first match between the test image and all images in the database, including the $K-th$ image, for all points not considered before the second match for the $K-th$ image has been established.

4.4 Image matching

In the previous section, the point matching procedure was introduced in order to determine the rotation and translation parameters. This procedure establishes the three point match between the test image and an image in the database - if such an image exists. To determine translation, x_0 and y_0, and rotation, θ, two matching points are enough. The unknown parameters may be calculated by using

$$x = x'cos\theta - y'sin\theta + x_0 \tag{4.15}$$

and

$$y = x'sin\theta + y'cos\theta + y_0, \tag{4.16}$$

where x and y are the test image coordinates and x' and y' are the database image coordinates. From three pairs of points obtained by the point matching procedure, three pairs of rotation/translation parameters are calculated. Then, the three unknowns, θ, x_0 and y_0 are obtained by averaging values from three computations. In this way, the computation accuracy is improved.

Calculated values θ, x_0, and y_0 are used in the image matching procedure to check for other points in the test image that have a match in the database

image labelled as a possible match. This is performed by using Equations (4.15) and (4.16) for every point in the test image. Ideally, if every point in the test image has a corresponding point in the labelled image, two images are matched. However, the present version of the algorithm assigns a match between the test image and the database image if the score

$$S = \frac{t_1 + 0.9t_2 + 0.8t_3}{t_t} \tag{4.17}$$

is greater than 0.9. Here,

- t_t is the total number of points in the test image
- t_1 is the number of points in the database image that match corresponding points in the test image within their 3×3 neighborhoods
- t_2 is the number of points in the database image that match corresponding test image points within their 5×5 and not within their 3×3 neighborhoods and
- t_3 is the number of points in the database image that match corresponding test image points within their 7×7 and not within their 5×5 neighborhoods.

If end point images match each other in the sense just explained, the last step in the authentication procedure is performed: fiber image matching by correlation. In this step, fiber images of the test document and a document chosen from the database are compared. If each fiber point in the test image has a corresponding point in the database image, the authentication procedure

is successfully completed. A *corresponding* point is considered in the 3×3 neighborhood of the point obtained by using Equations (4.2) and (4.3). Otherwise, the authentication procedure continues until either all database images are examined or a match is established.

CHAPTER 5

Probabilities of Uniqueness

This chapter analyzes the uniqueness of documents containing fibers and the uniqueness of documents containing fiber end points.

Two different fiber shapes are considered: (1) line segments and (2) arbitrary shapes. In both cases we evaluate the uniqueness of documents containing fibers and the uniqueness of documents containing fiber end points.

5.1 Case 1: Line segments

5.1.1 Fiber case

Suppose that a document consists of fibers that can adequately be modeled as line segments. We estimate the number of different documents that can be formed as a function of fiber density (fiber density is specified as number of fibers per square inch).

Let us denote K_1 the number of possible combinations of documents and K the number of detectable combinations of documents that may be formed. If documents are not densely covered by fibers (small d), $K \approx K_1$. For high densities, the saturation effect is significant, and K becomes less than K_1. (A new fiber placed in a region already covered by fibers does not carry any new

information.) If K_2 represents the saturation effect, the number of detectable combinations is then $K = K_1 - K_2$.

Next, both K_1 and K_2 are estimated.

Estimation of K_1

Let us assume the following:

- document size: 8.5×11 $[inch^2]$
- fiber density: $d[fibers/inch^2]$
- pixel size: $0.5 \times 0.5[mm^2]$ and
- fiber length: $10[pixels]$.

The number of pixels per document and the number of fibers per document are $m = 241,289$ and $n = 8.5 \times 11 \times d$, respectively. For a single fiber, the number of different combinations in a document is $N = s_1 * m = 3,378,046$, where $s1$ is the number of different combinations when the central point of the fiber is fixed (under given assumptions $s_1 = 14$). Therefore, K_1 is given as

$$K_1 = \begin{pmatrix} N \\ n \end{pmatrix} \tag{5.1}$$

and is a function of fiber density, d. Note that edge effects are not taken into account where N is calculated. Since the fiber length is much less than the document size, edge effects may be neglected.

Estimation of K_2

Suppose that there are n fibers in a document which occupy $m_1 = n \times 10$ pixels. (This is the worst case because fiber overlappings are not taken into consideration.) Suppose, also, that all fibers are placed into a region without empty pixels; i.e., all pixels in that region are covered by fibers. (This is the worst case, also, because it gives the largest possible value for K_2.) If d is the fiber density in the document, then

- for a single fiber, the number of different combinations in the depletion region (region occupied by fibers) is given by $N_1 = m_1 \times s_1$
- the number of fibers that belong to the depleted region is given by $n_1 = n\ N_1/N$.

Consequently, K_2 is given by

$$K_2 = \begin{pmatrix} N_1 \\ \\ n_1 \end{pmatrix}, \tag{5.2}$$

and this number is a function of fiber density d.

Estimation of the number of different combinations

Suppose that numbers K_1 and K_2 are obtained as shown by Equation (5.1) and (5.2). Then, the number of different combinations of documents is calculated as $K = K_1 - K_2$.

When fiber densities are small, K_2 can be neglected, and thus, $K \approx K_1$. However, when fiber density increases, K_2 increases also. For large densities,

K_1 and K_2 become very large numbers and $K_1 - K_2$ is of the type $\infty - \infty$. To gain some insight into the manner of how K changes with fiber density d, the expression $\frac{K_1}{K_2}$ is considered. When we take the logarithm of $\frac{K_1}{K_2}$, the fraction becomes $k_1 - k_2$, where $k_1 = log\ K_1$ and $k_2 = log\ K_2$. In Figure 5.1, $k_1 - k_2$ versus density d is shown. As expected, first $k_1 - k_2$ increases with d; then the increase becomes slower, and finally $k_1 - k_2$ starts to decrease when d continues to increase. The saturation effect is significant in regions with negative slope as in Figure 5.1.

5.1.2 End point case

The probability of uniqueness for the end point case is determined by Equations (5.1) and (5.2) because fiber length is constant. Thus, uniqueness for the fiber case and that for the end point case are the same when fibers are line segments. Strictly speaking, uniqueness for the end point case is slightly less than that established for the fiber case (Equations (5.1) and (5.2)) because of the fact that some situations, like total fiber overlappings or end point overlappings, are not detectable by the end point detection algorithm. Because such cases are fairly rare, corrections are not considered.

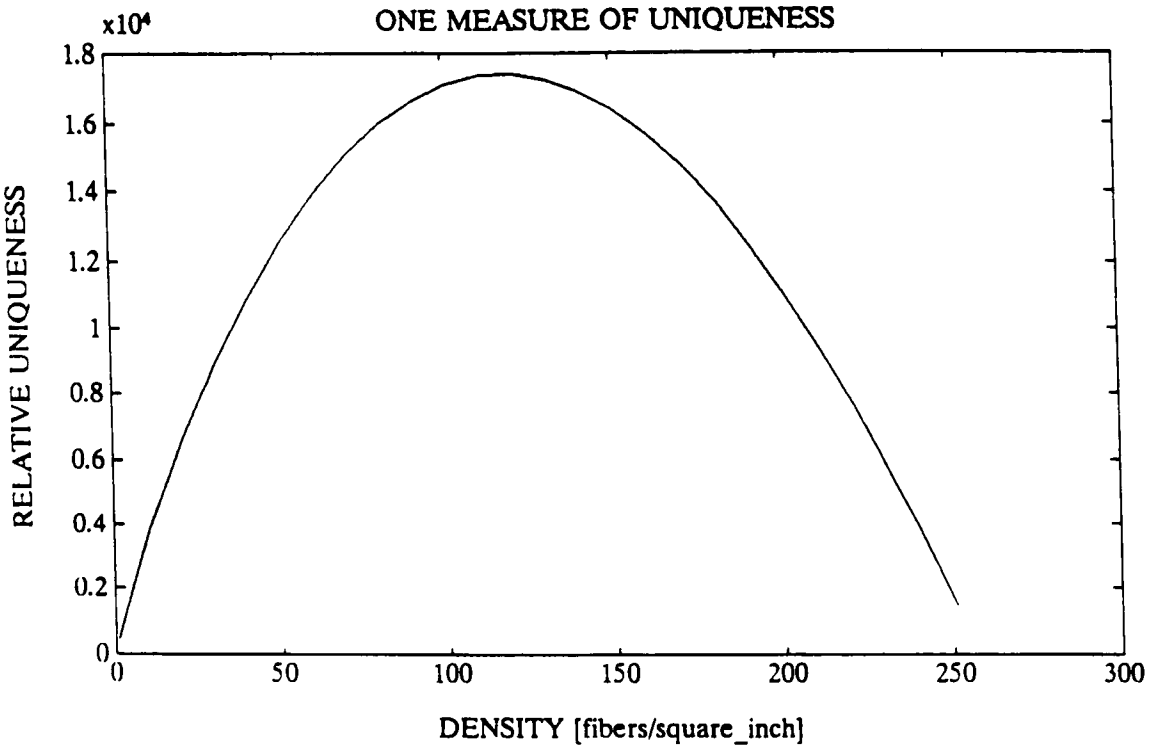


Figure 5.1: Uniqueness as a function of fiber density for fibers as line segments.

5.2 Case 2: Fibers of arbitrary shapes

5.2.1 Fiber case

We estimate the probability of uniqueness for documents containing fibers that are not necessarily line segments, i.e., documents containing fibers of arbitrary shapes. Arbitrary fiber shape means that

- the first fiber point (e.g., one of the fiber end points) is somewhere in the middle of the document
- the second fiber point is one of the eight neighbouring points
- the third, fourth, and so on until the tenth fiber point (the last fiber point) is one of the next 5 closest neighbouring points in the same direction.

Thus, for a fixed point in a document (i.e., is the first fiber point), a fiber may have $s_2 = 8 \times 5^8 = 3,125,000$ different possible shapes. So, to calculate K_1 and K_2 or $k_1 - k_2$ for documents consisting of fibers of arbitrary shapes, s_1 is to be replaced by s_2 in Equations (5.1) and (5.2). Everything else remains unchanged. In Figure 5.2, $k_1 - k_2$ versus density d is shown. As we see, the shapes of the curves $k_1 - k_2$ versus d for two cases (fibers are line segments and fibers of arbitrary shapes) are alike (see Figure 5.1).

5.2.2 End point case

In this section the probability of uniqueness of a document consisting only of fiber end points will be evaluated. Arbitrary fiber shape is assumed.

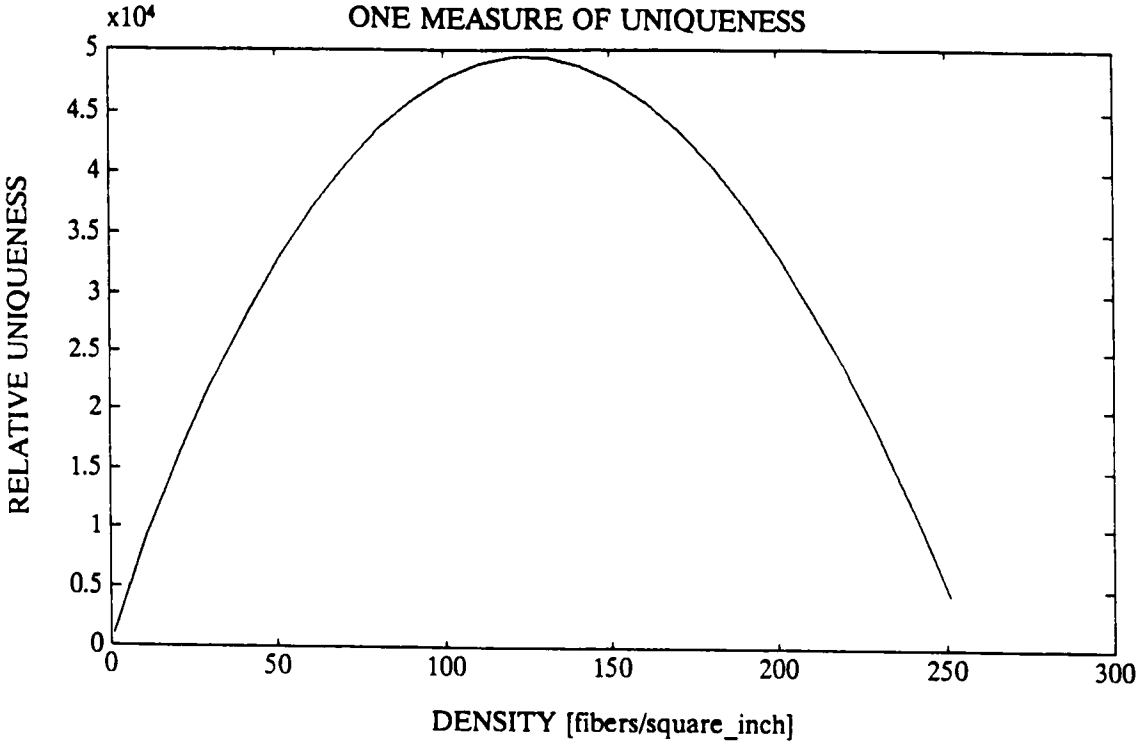


Figure 5.2: Uniqueness as a function of fiber density for documents consisting of fibers of arbitrary shapes.

Basic assumptions related to document size, pixel size, and fiber length remain unchanged. Besides those assumptions, the following calculation is performed considering:

- maximum and minimum distance between fiber end points of 10 and 8 pixels, respectively and
- digitization effect of 1 pixel.

Suppose that there is a fiber in a document. If the first pixel of the fiber is placed somewhere in the middle of the document, the last fiber point may be located at 6, 7, 8, 9, 10, 11 pixels away from the fiber's first point. (Note that the minimum and the maximum distance between the end points of a fiber are combined with the digitization effect.) This means that the equivalent, new pixel size is 6 pixels. In other words, the document size, as well as the fiber length, is to be compressed. Since the fiber length is 10 pixels, the equivalent fiber length is 1 new, equivalent pixel. (Actually, the length is either 1 or 2 , but we consider 1 only because it is the worst case.) At the same time, a document consists of $m^* = 6603$ new pixels instead of $m = 241289$ points like before. The number of different combinations for a fiber in a document is $N^* = 6603$ and the number of different documents is $\begin{pmatrix} 6603 \\ n \end{pmatrix}$, where n denotes the number of fibers in a document.

In Figure 5.3, number of combinations versus fiber density is shown for the end point case if fibers are of arbitrary shapes. Note that the shapes of curves in Figures 5.1, 5.2, and 5.3 are alike as is expected. All these curves differ

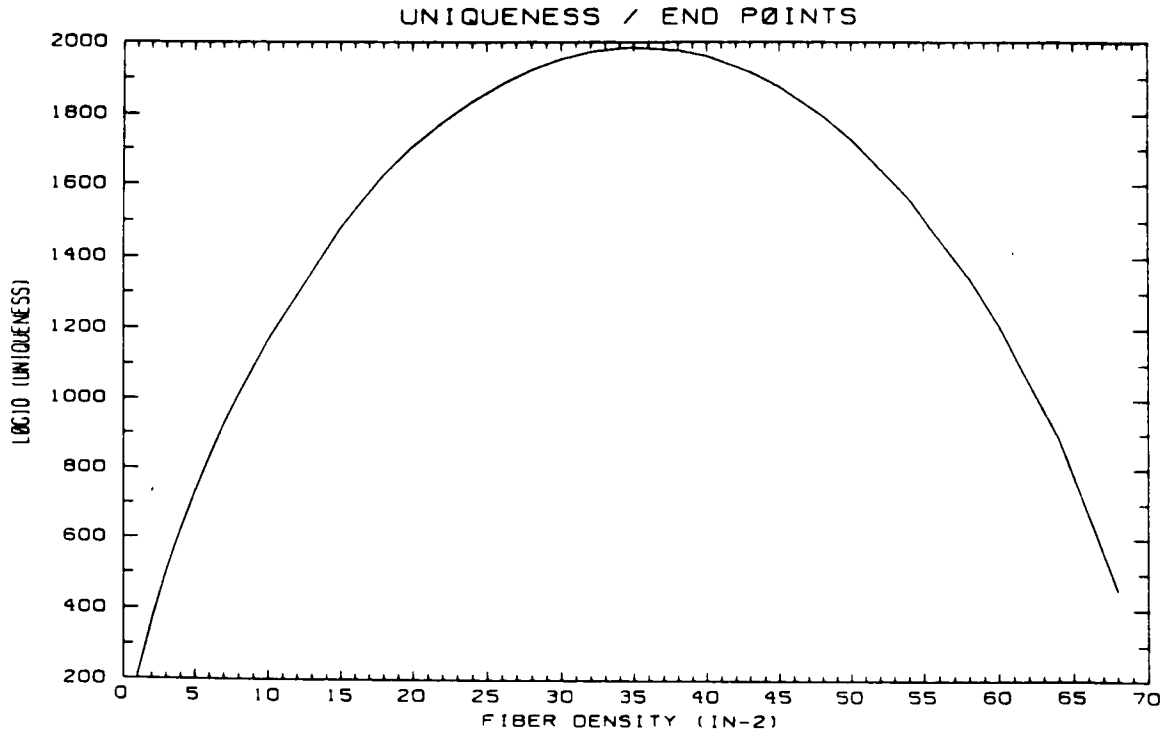


Figure 5.3: Uniqueness of the end point case: fibers are of arbitrary shapes.

among themselves by constant factors. Also, horizontal shifts between them are possible.

CHAPTER 6

An Alternative Approach to Document Authentication

This section introduces a new approach in document authentication: assigning a unique code to each document. Three different concepts based on distances between fiber end points are considered. The first concept is based on the histogram of distances between end points. The second is based on an ascending order of distances between all end points in an image. The third one is based on distances as a function of phase angles between end points in an image and a prespecified reference point.

There are several reasons why image encoding is considered in this thesis.

- First, the database contains a large number of images of size 256×256. Such a database requires a large memory space in a computer. Also, any search conducted throughout such a database is not convenient.

- Second, the document authentication procedure (explained in the previous section) becomes very slow when there is a large number of images in the database. This is caused by the image reading operation, which is very slow because a large amount of data must be read; for example, an image of the size 256×256 pixels occupies 132 memory blocks in a computer. As a matter of fact, most of the computer CPU time is consumed by the

reading operation while the document authentication is in progress.

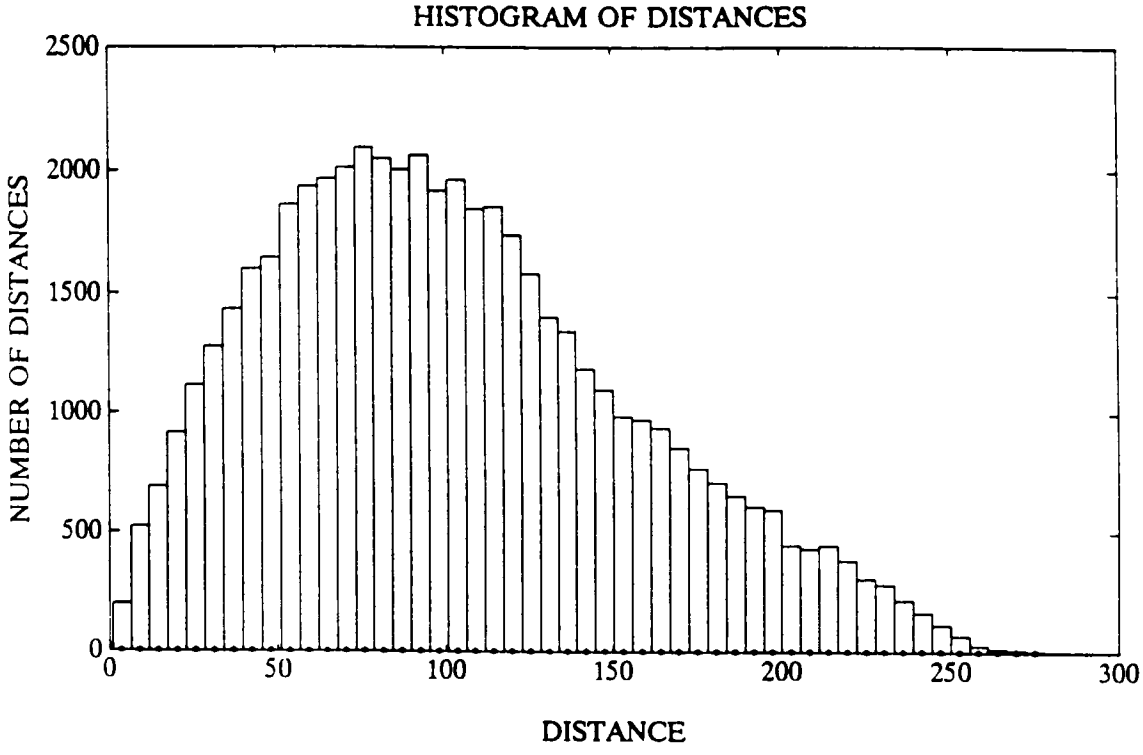
- Third, we want to assign a unique signature to an image so that the image can still be identified. Since a signature is a numerical code, the database can be organized not as a set of images but as a set of signatures according to their numerical values. Given a test image, its signature is compared to signatures in the database. In the worst case, the test image and a small subset of database images are to be compared.

An image is subject to rotation and/or translation as well as to a fiber removal process. So, the objective of image encoding is to create a signature that is invariant to rotation and translation. Also, the signature should not change (or should change as is predetermined) when some fibers are removed from the original document.

6.1 An approach based on histogram of distances

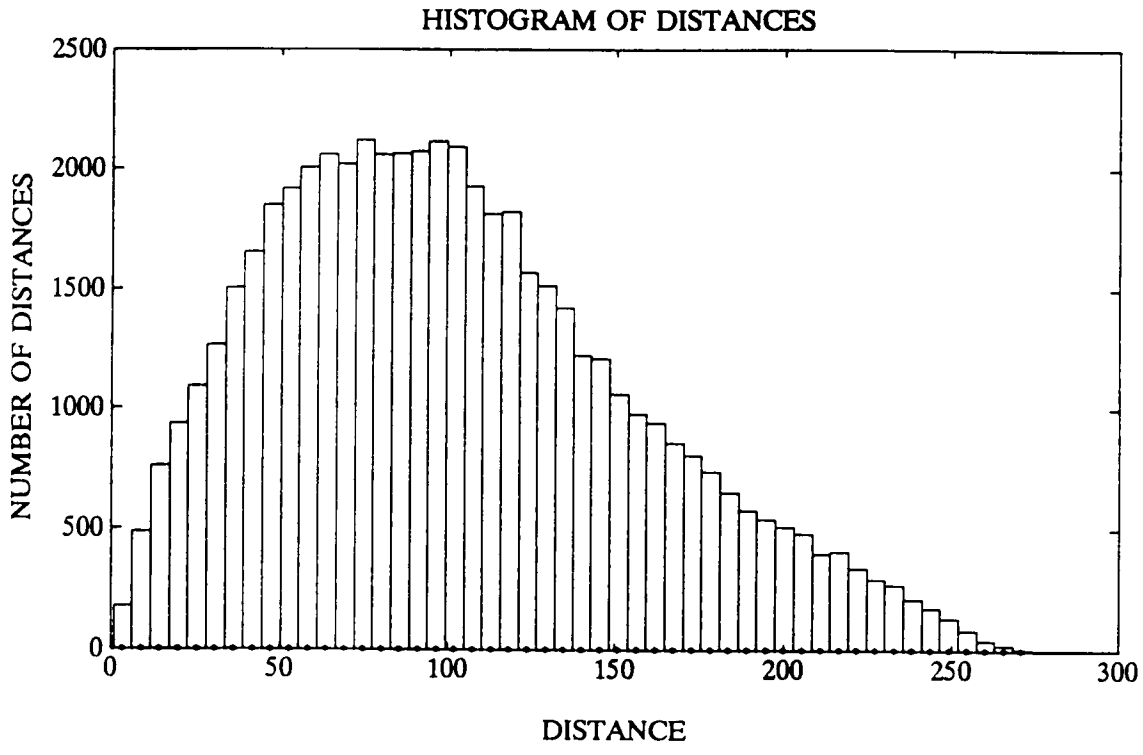
Suppose that an arbitrary document and its end point image are given. A histogram of distances between points in the end point image is shown in Figure 6.1 (a). In Figure 6.1 (b), a histogram of distances between the fiber end points of another document is shown.

From Figure 6.1 we see that distances are not uniformly distributed; in fact, they approach Gaussian or Rayleigh distribution. Equivalently, all distances in the image do not carry the same amount of information, and thus, all distances are not equally *important*. Hence, the numerical code resulting from the his-



(a)

Figure 6.1: Histograms of distances of two different images.



(b)

Figure 6.1: Continued.

togram of distances should be weighted in some manner so that more *important* distances get larger weights (for example, distances less than 50 and greater than 300) and less *important* distances get smaller weights (for example, distances between 50 and 300).

There are several advantages to document coding based on the histogram of distances approach:

- Histograms are invariant to rotation and translation, which is not surprising since distances *among* fiber end points are considered (not distances from end points to prespecified reference points that may be changed when image acquisition is performed).

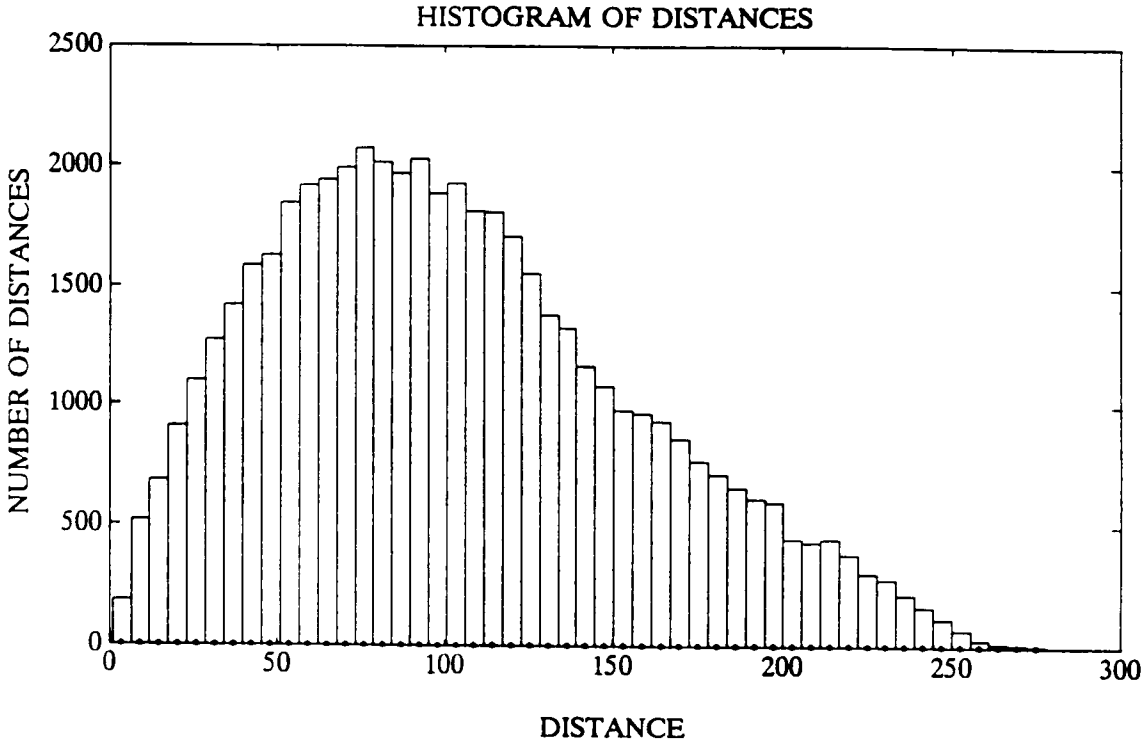
- Missing fibers do not change histograms significantly. In Figure 6.2, several histograms of an image with one, two, and twenty missing fibers are shown; a histogram of the original document is shown in Figure 6.1 (a).

- Missing fibers do not affect minimum and maximum distance. Compare Figures 6.1 (a) and Figures 6.2 (a) through (c).

Disadvantages:

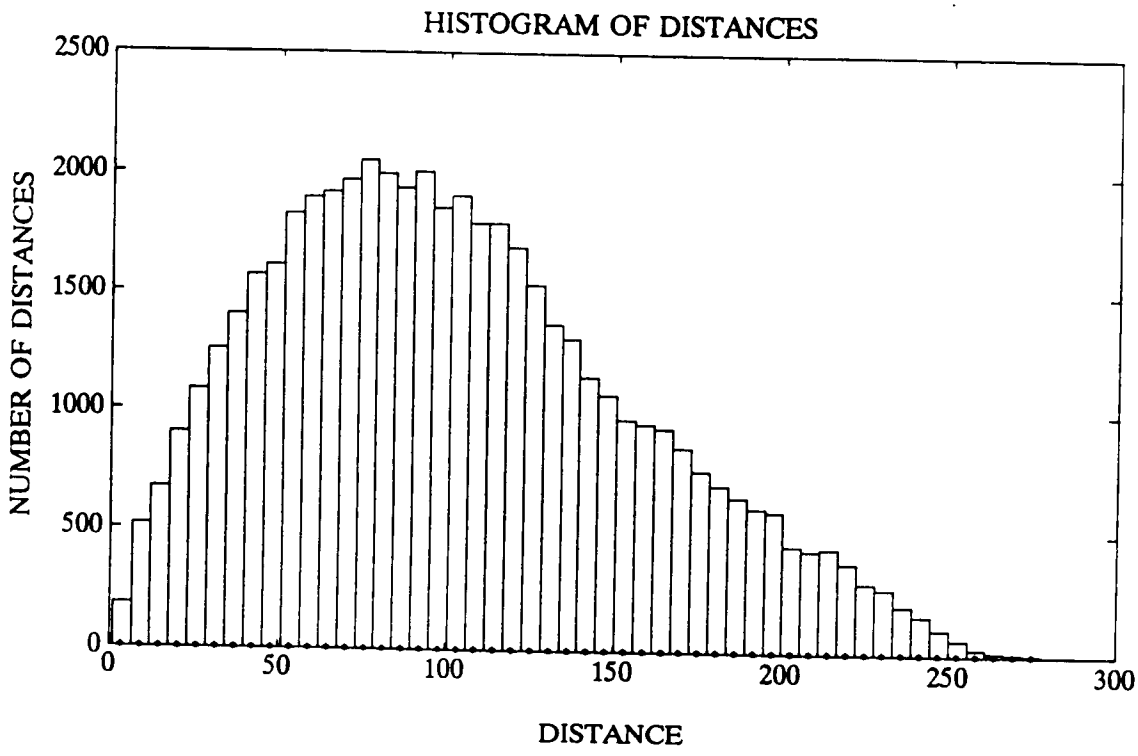
- It is not clear how to choose weight factors for distances so that resulting codes have maximum probability of uniqueness.

- Histograms of different images are alike and differ slightly among themselves (compare figures (a) and (b), Figure 6.1).



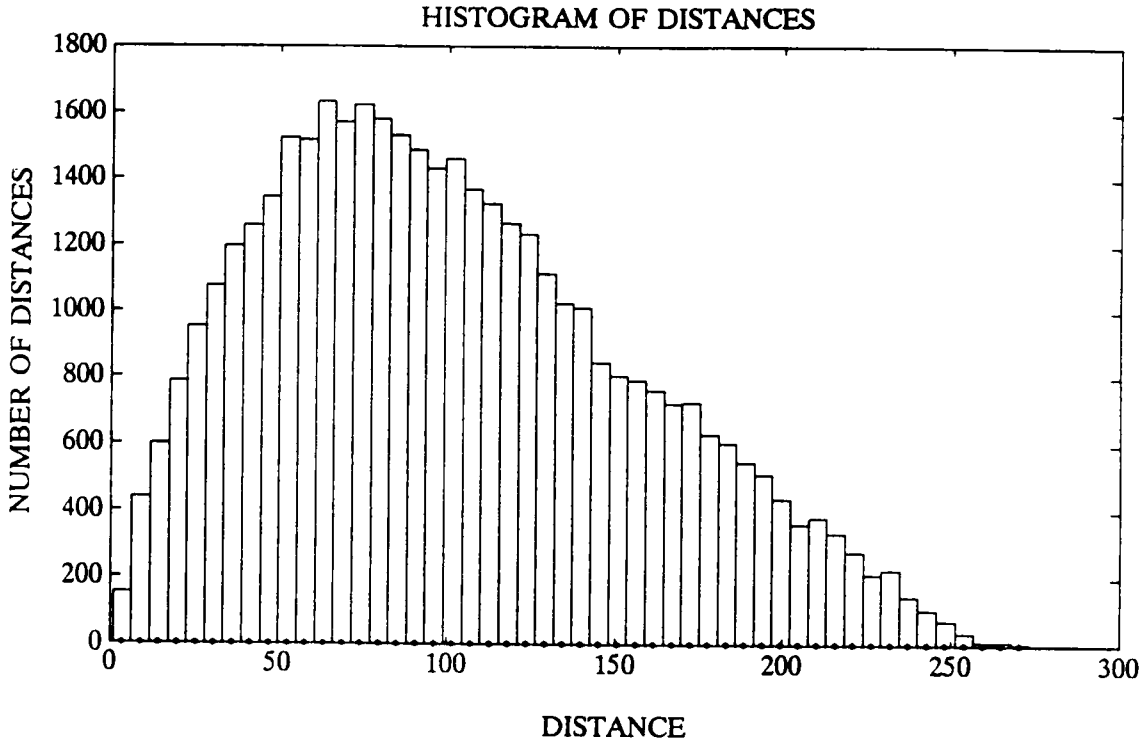
(a)

Figure 6.2: Histograms of an image with (a) one missing fiber, (b) two missing fibers, and (c) twenty missing fibers.



(b)

Figure 6.2: Continued.



(c)

Figure 6.2: Continued.

6.2 An approach based on an ascending order of distances

This approach in image coding is based on an ascending order of distances between fiber end points in documents. Once the end point image of an arbitrary document is obtained, all distances between end points are determined. Then, distances are sorted in an ascending order. For an arbitrary document, ascending order of distances between fiber end points is shown in Figure 6.3.

The main advantage of this approach is that graphs of distances, like the one shown in Figure 6.3, are invariant to translation and rotation. Also, the largest and the smallest distances in graphs remain unchanged when some fibers in documents are missing. (See Figure 6.4 where ascending order of distances is shown for a document when (a) one, (b) two, and (c) twenty fibers are missing, respectively; ascending order of distances of the original document is drawn by solid lines, while those of images with missing fibers are drawn by dashed lines).

However, there are some disadvantages to this approach:

- Codes become very large if no information is lost (about $50,000$ numbers each).
- If some fibers are missing, the graph of ascending order of distances is changed significantly (Figure 6.4 (a) through (c) shows how the graph of a document is changed when (a) one, (b) two, and (c) twenty fibers are missing, respectively).
- The ascending orders of distances of two different documents are alike;

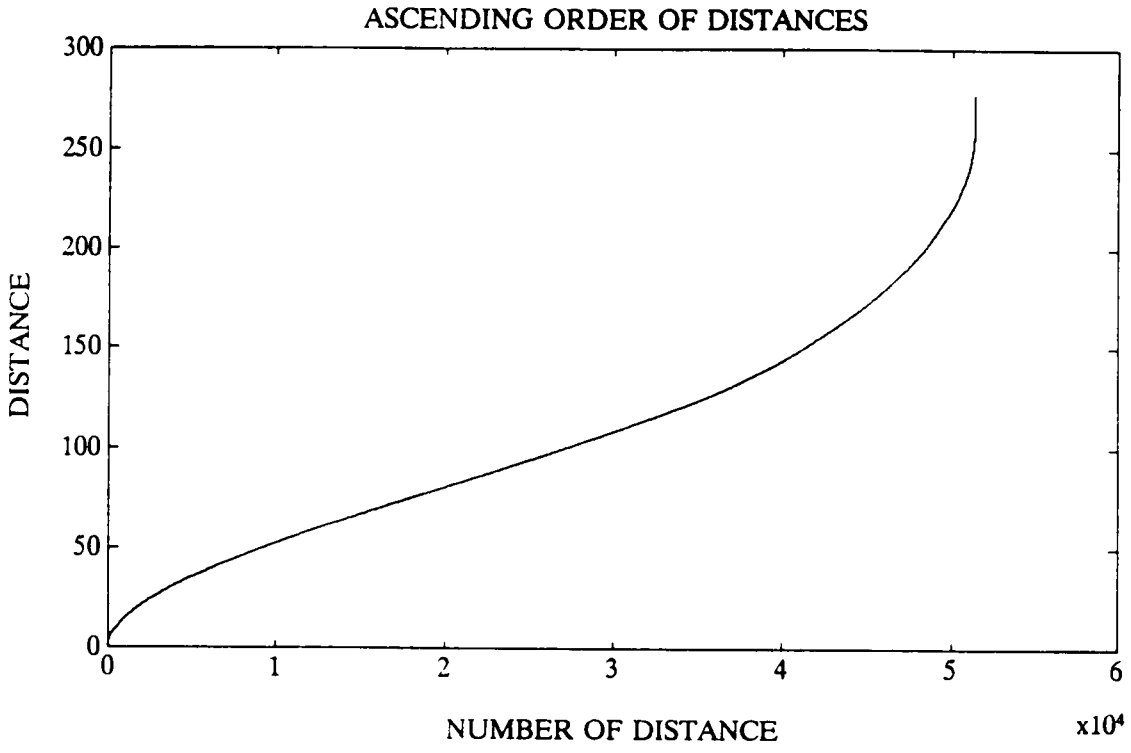
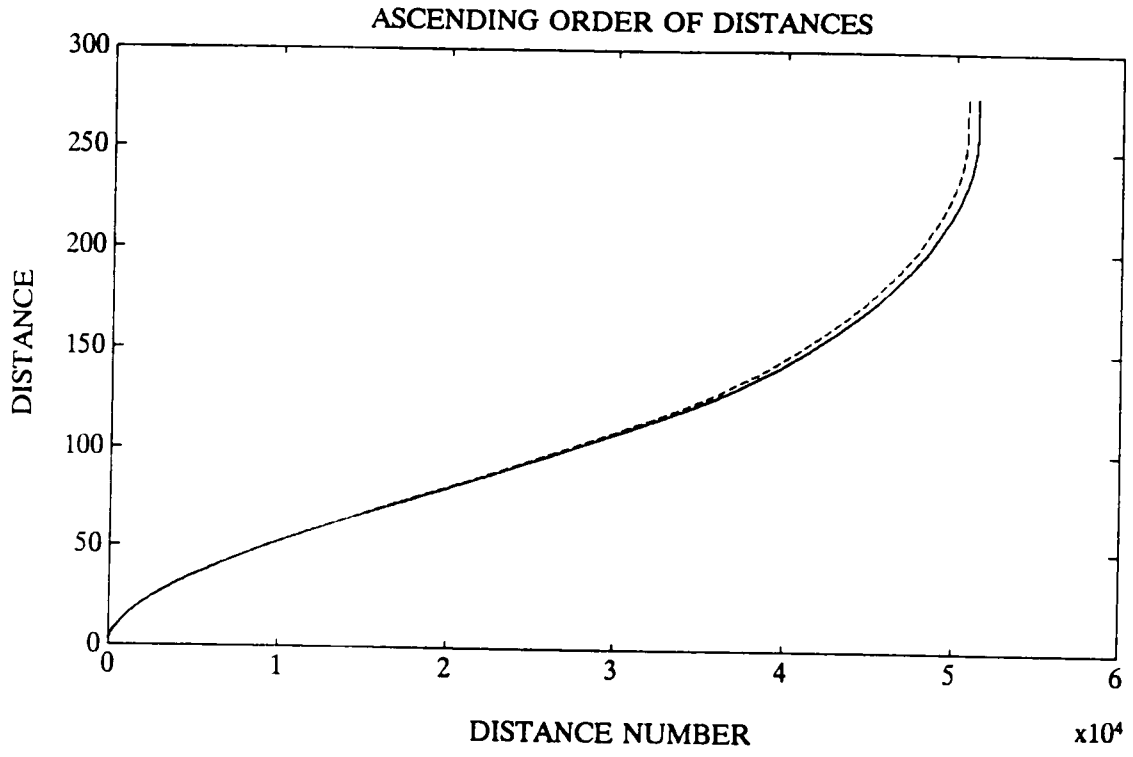
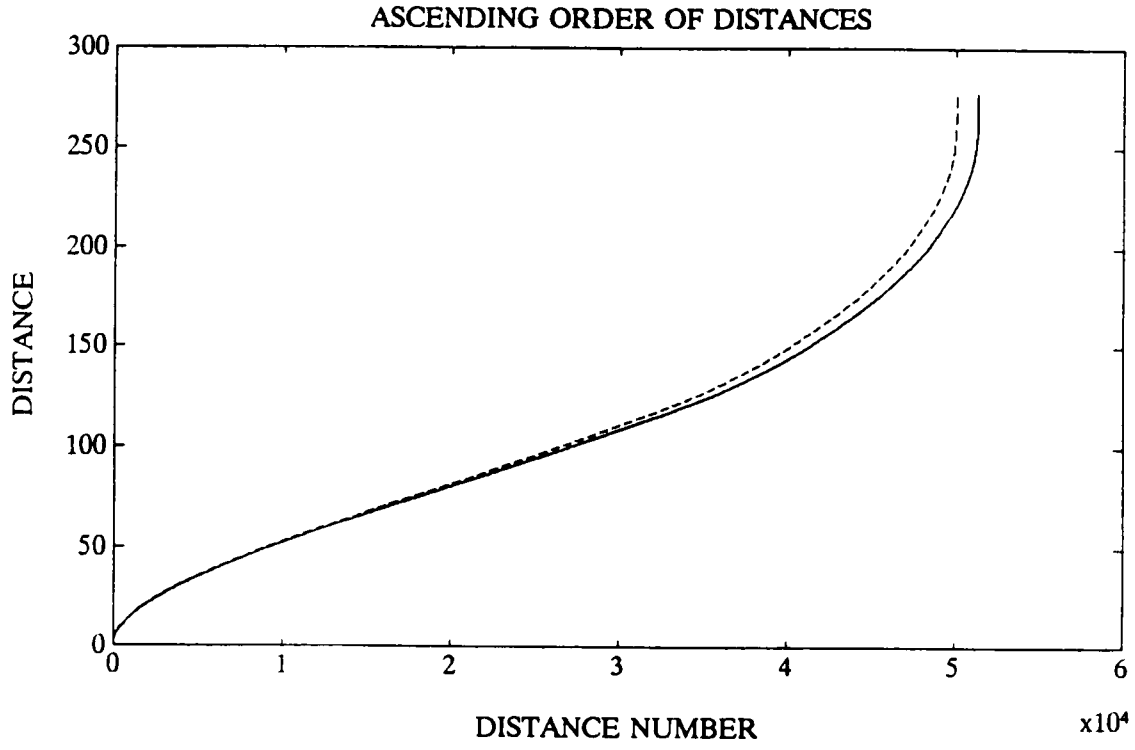


Figure 6.3: Ascending order of distances of an arbitrary document.



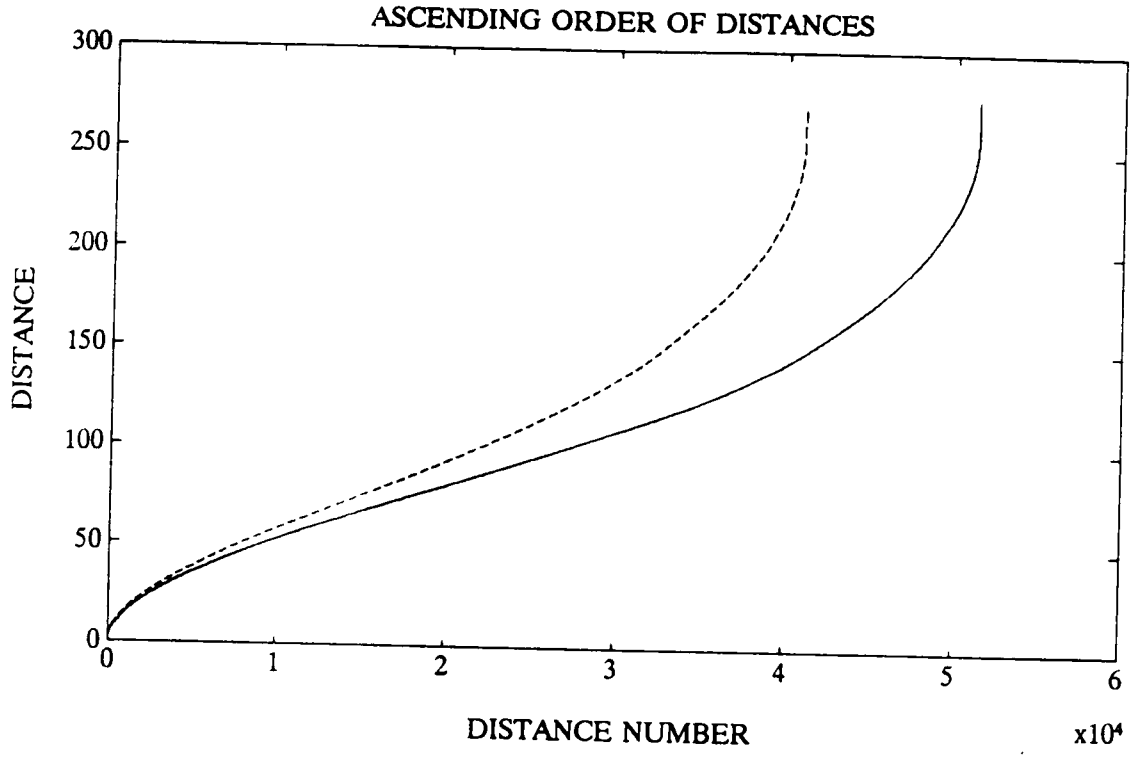
(a)

Figure 6.4: Ascending order of distances of a document with (a) one, (b) two, and (c) twenty fibers missing (dashed lines); ascending order of distances of the original document is shown by solid lines.



(b)

Figure 6.4: Continued.



(c)

Figure 6.4: Continued.

this may affect code uniqueness when a large number of documents is to be encoded (see Figure 6.5).

6.3 An approach based on distances and phase angles

Suppose that the end point image of an arbitray document is given. Suppose also that a specific point in the end point image is chosen as a reference point. Then, distances, r, and phase angles, θ, between image points and the reference point are found and graph $r(\theta)$ is formed. Finally, the code based on the graph $r(\theta)$ is obtained. In Figure 6.6, graphs $r(\theta)$ of two different images are shown.

Advantages of the concept based on distances and phase angles include:

- Graphs $r(\theta)$ of the test and the original image are the same when two images are translated relative to each other.

- Graphs $r(\theta)$ of different images are significantly different among themselves. (If the difference is not obvious from Figures 6.6 (a) and (b) , see Figures 6.7 and 6.8; in Figure 6.7 histograms of graphs in Figure 6.6 are shown, while in Figure 6.8 the difference between histograms in Figure 6.7 is shown.) Because graphs $r(\theta)$ are different for different images, uniqueness of codes based on these graphs is large compared to uniqueness of codes from the previous method.

- Missing fibers do not affect graphs $r(\theta)$ significantly. Compare graphs

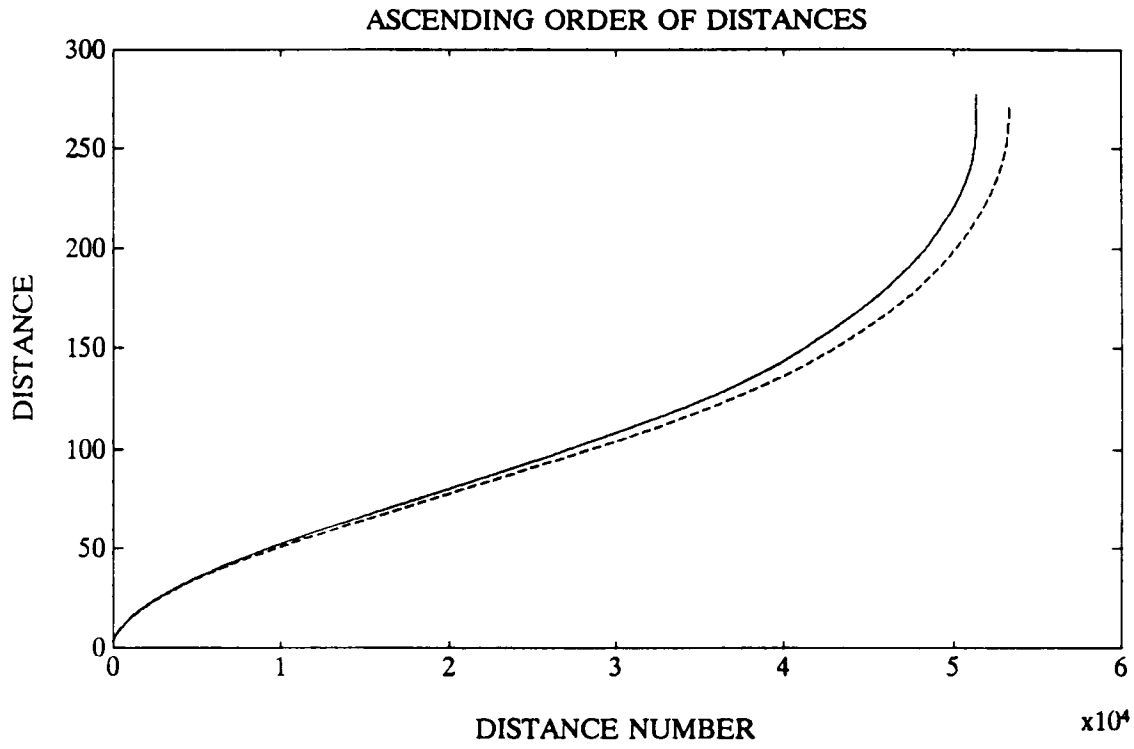
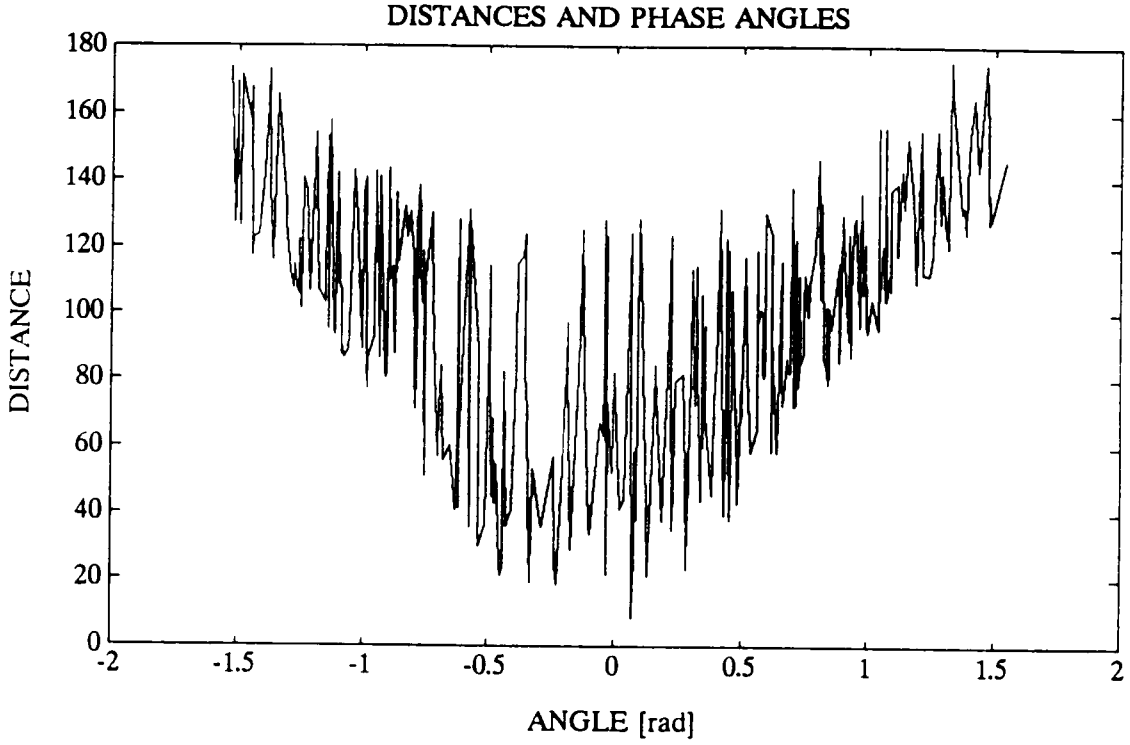
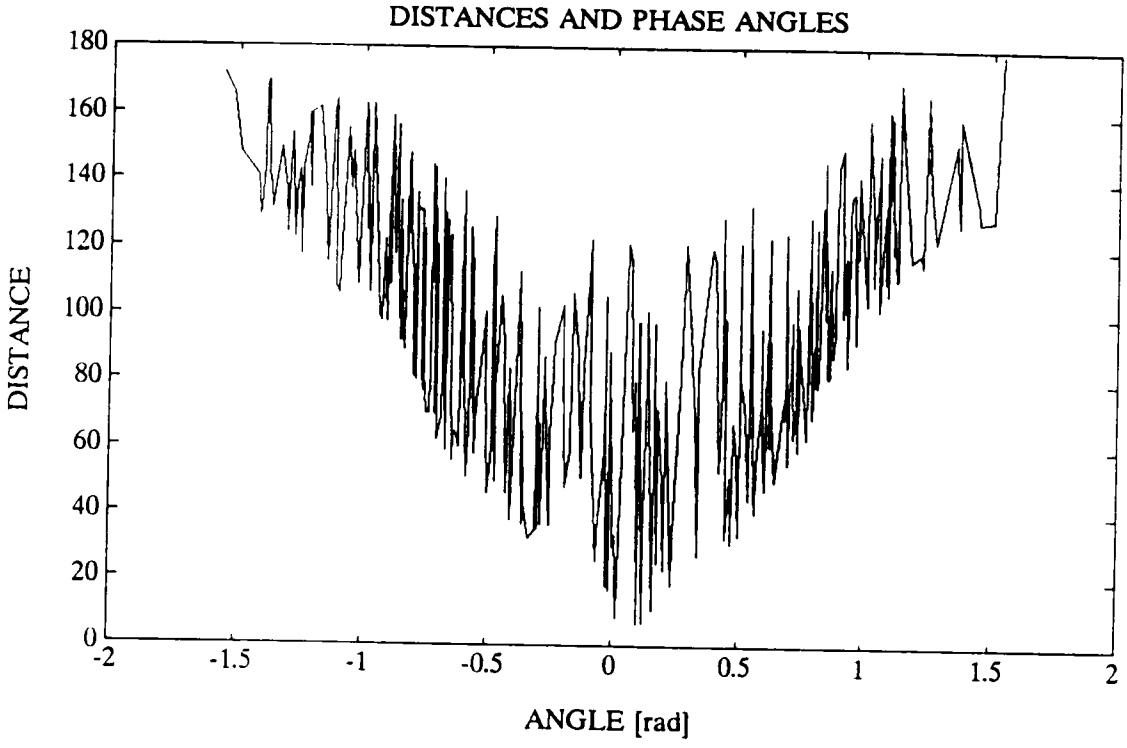


Figure 6.5: Ascending order of distances of two different documents.



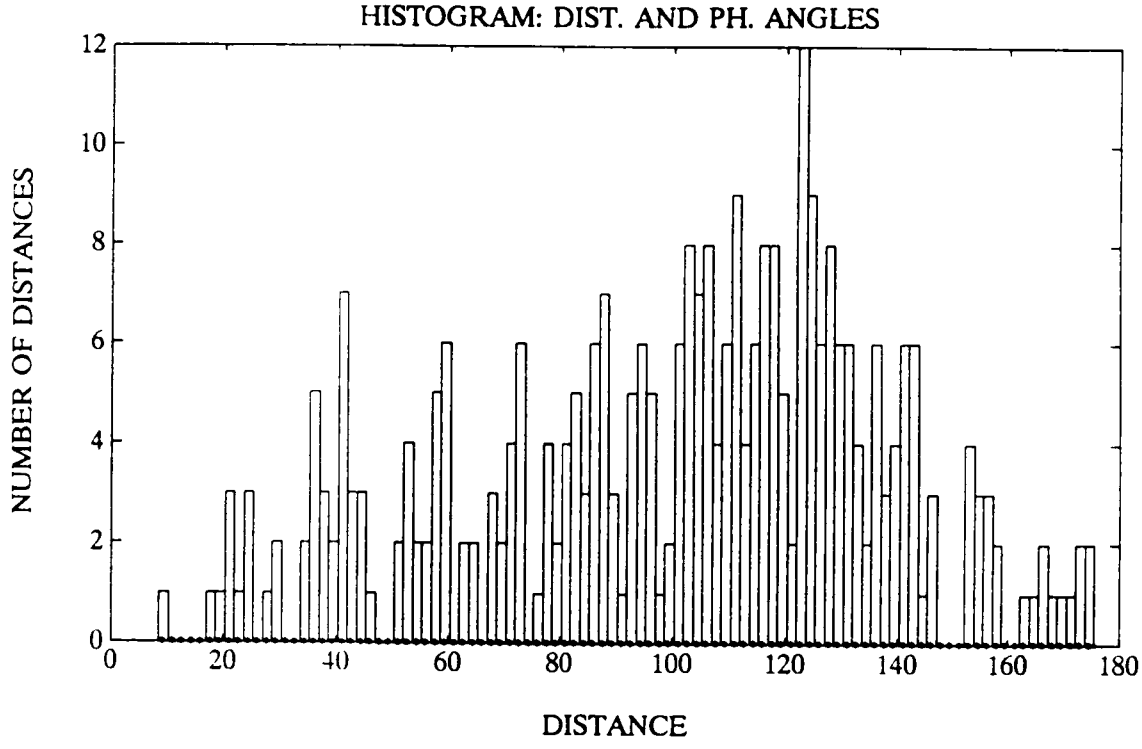
(a)

Figure 6.6: Graphs $r(\theta)$ for two different documents.



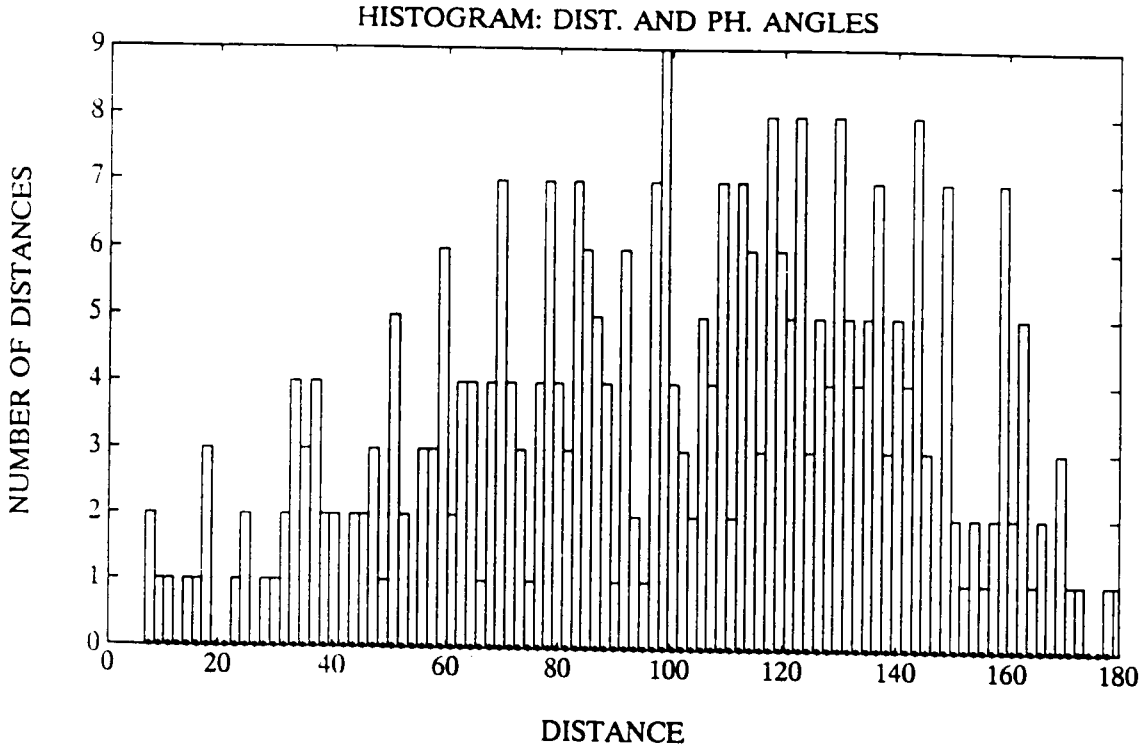
(b)

Figure 6.6: Continued.



(a)

Figure 6.7: Histograms of graphs $r(\theta)$ from Figure 6.6.



(b)

Figure 6.7: Continued.

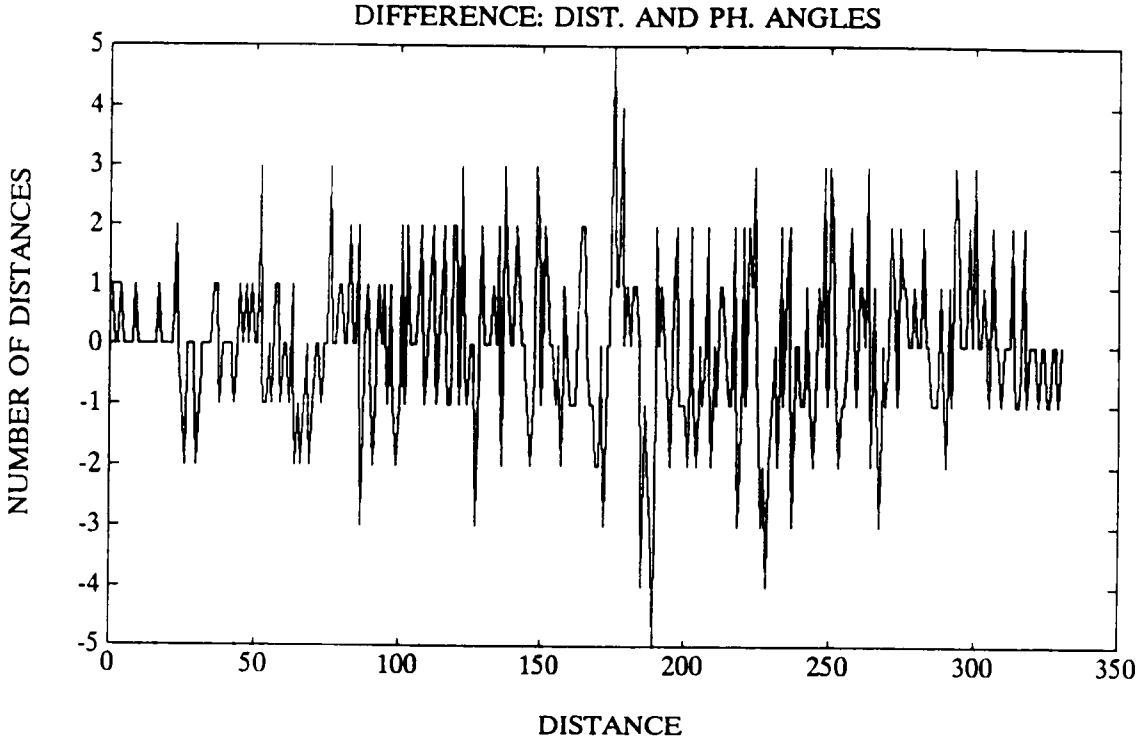


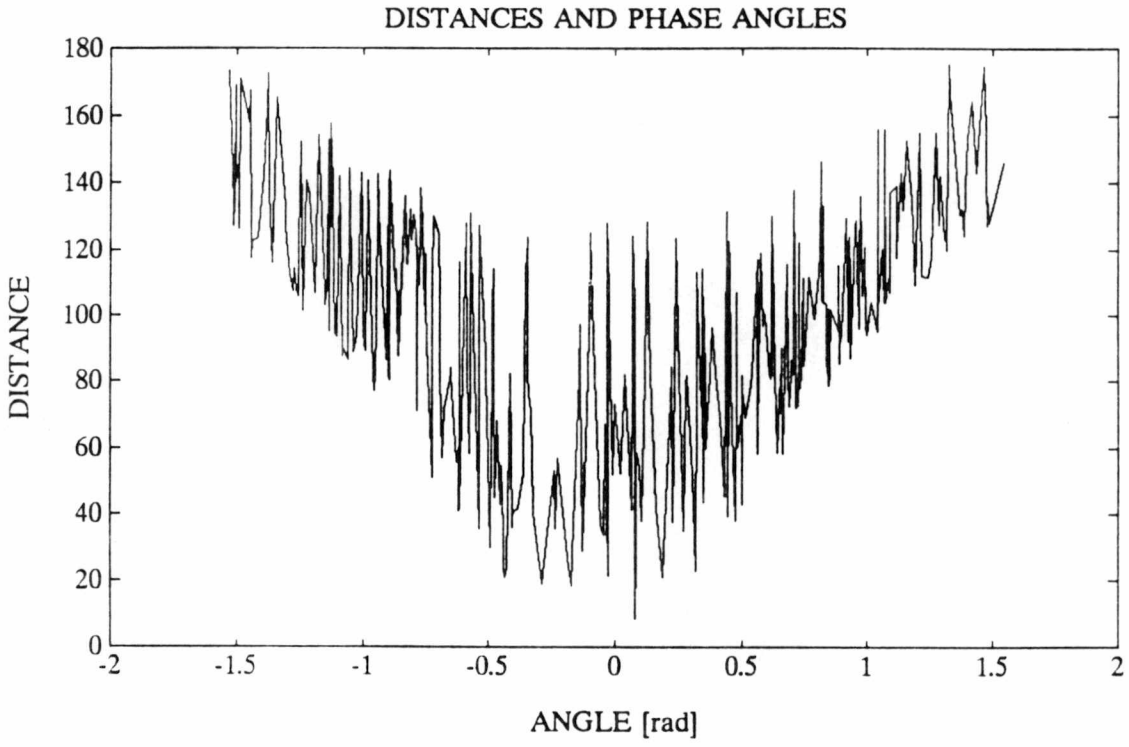
Figure 6.8: Difference between histograms (a) and (b) from Figure 6.7.

$r(\theta)$ of the original document, Figure 6.6 (a), and graphs $r(\theta)$ of the same document with one, two, and twenty fibers missing, as shown in Figures 6.9 (a) through (c); corresponding histograms are shown in Figure 6.10.

- Codes based on graphs $r(\theta)$ contain about 600 numbers (compare to $50,000$ required for codes in the previous approach).

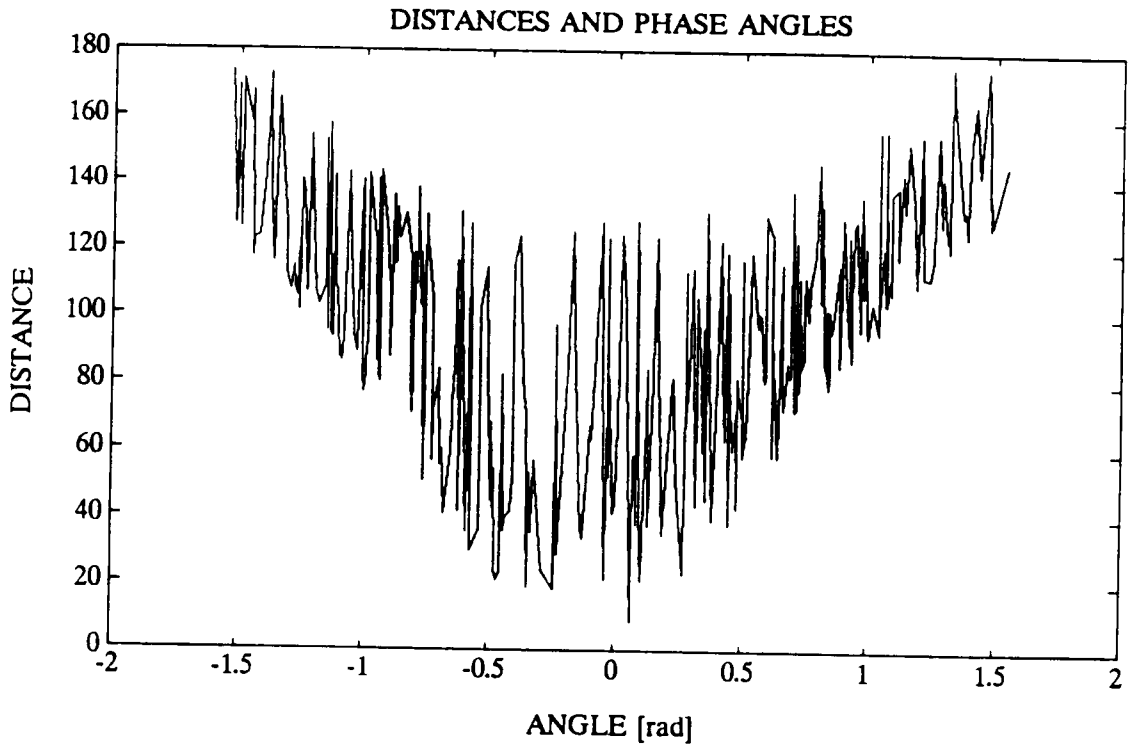
Disadvantages:

- Criteria for choosing referent point are ambiguous.
- When the original document and the test document are rotated relative to each other, graphs $r(\theta)$ of two images are shifted, depending on the rotation angle, but the shape of the graphs are the same; hence, graphs are not invariant to rotation, unlike the previous two approaches.



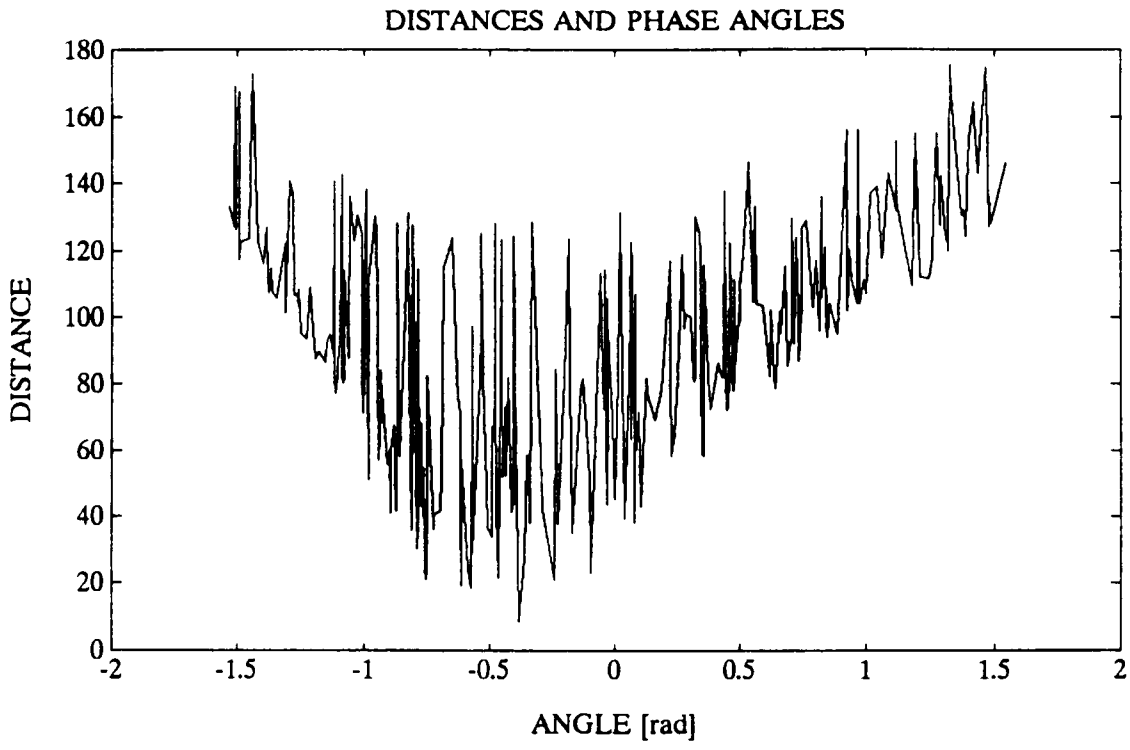
(a)

Figure 6.9: Graph $r(\theta)$ from Figure 6.6 (a) when (a) one, (b) two, and (c) twenty fibers are missing.



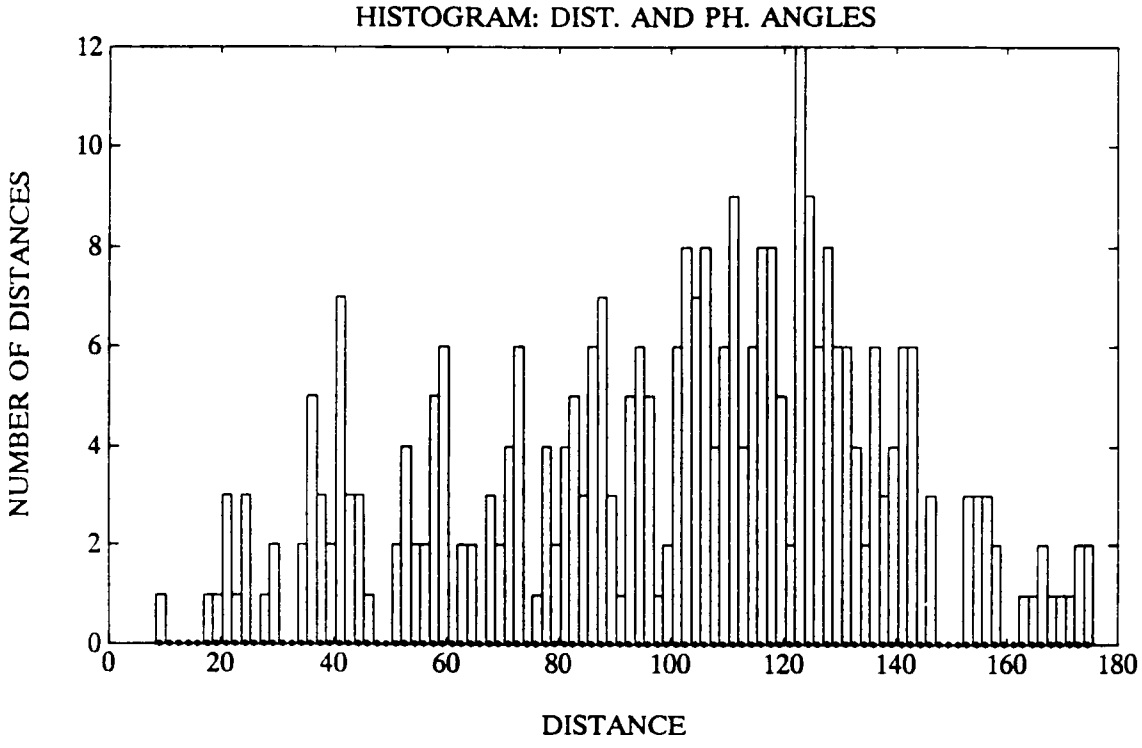
(b)

Figure 6.9: Continued.



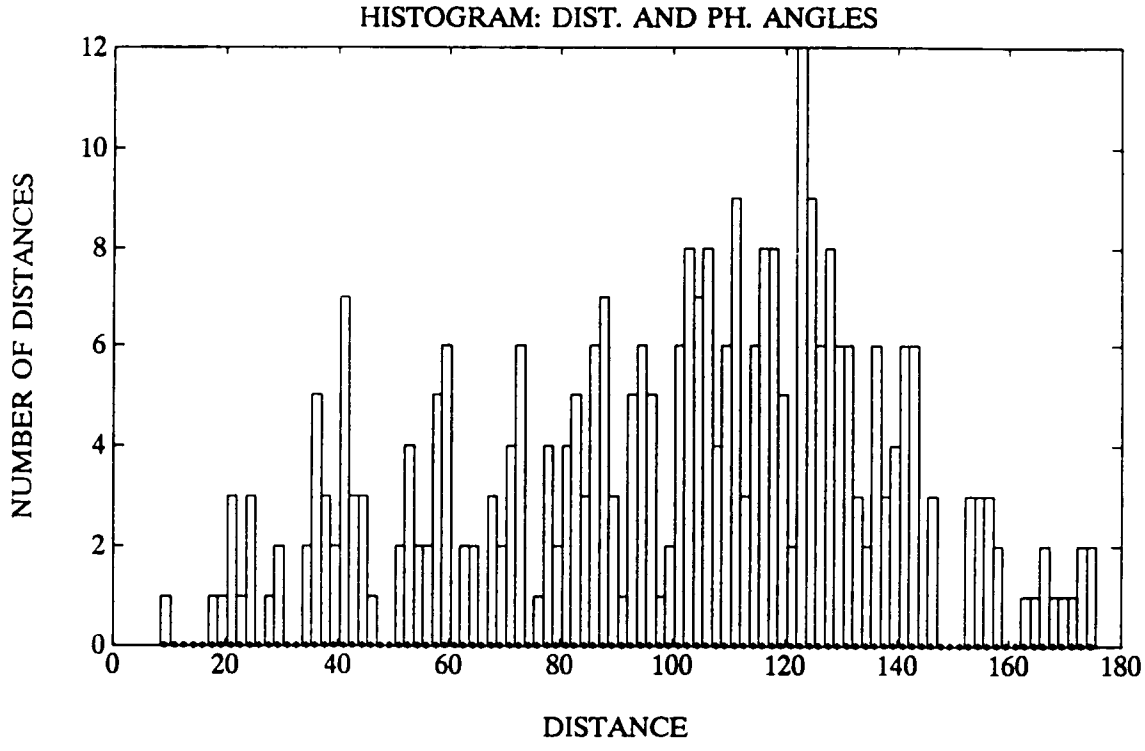
(c)

Figure 6.9: Continued.



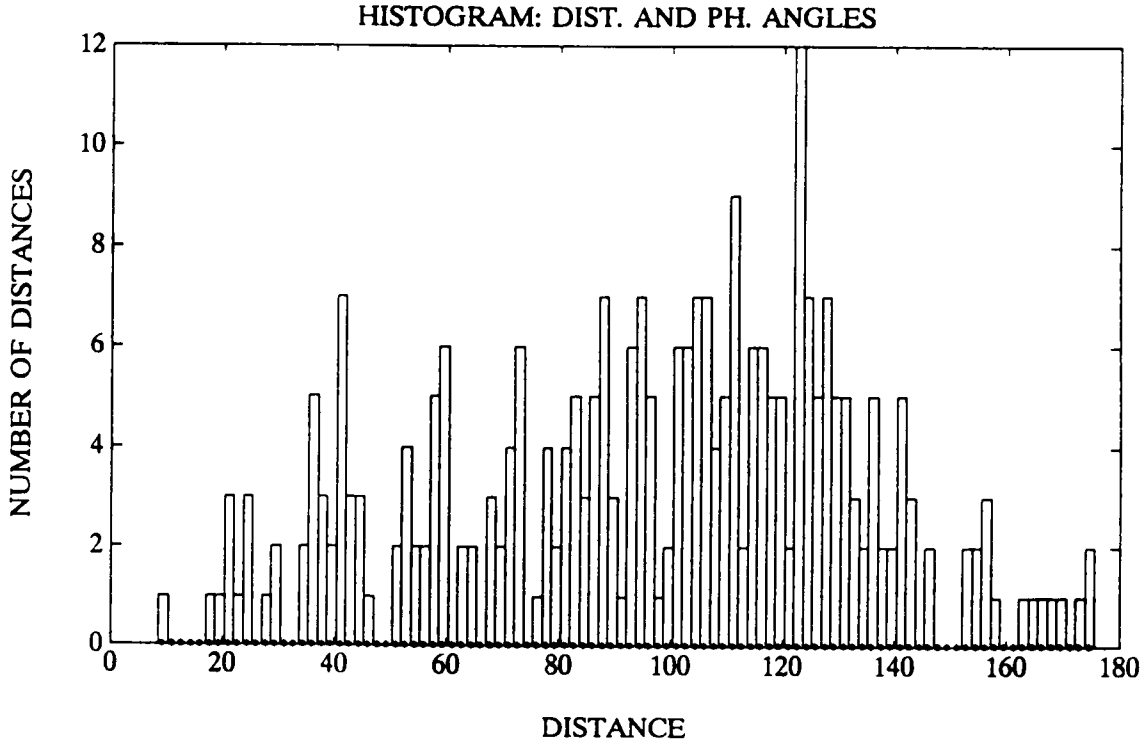
(a)

Figure 6.10: Histograms from Figure 6.9.



(b)

Figure 6.10: Continued.



(c)

Figure 6.10: Continued.

CHAPTER 7

Conclusions

In this thesis a new approach to the document authentication problem is introduced. Given a set of documents representing a database and a test document, we want to determine whether or not the test document belongs to the database. Documents that are considered in the thesis consist of optical fibers randomly scattered onto noisy backgrounds. The test document is subject to translation and rotation, and it may contain fewer fibers relative to the database documents. The database contains a large number of different documents. There are no control points or predetermined features in either the test or the database documents. Thus, there are two new assumptions in the document authentication process considered in this thesis that have not been considered before in most similar problems of registration and correspondence and they are (1) the database consists of a large number of different documents and (2) there are no control points in enther the test or the database images.

The present implementation of document authentication is guaranteed to work under the following assumptions:

(1) the maximum translation in the x direction is 20 pixels

(2) the maximum translation in the y direction is 20 pixels

(3) the maximum rotation angle is 3^0

(4) the maximum number of missing fibers randomly scattered over a document is 11 fibers

(5) the maximum of a fourth of a continual region containing fibers may be missing. The developed algorithm has been tested on more than a thousand different documents so far, all simulated on the computer. It was shown that the algorithm gave correct results wherever translation, rotation, and number of missing fibers were under specified limits ((1) through (5)). Actually, the algorithm may provide a correct answer if a test image is translated and rotated beyond the specified limits or if the number of missing fibers exceeds 11.

In addition to the above algorithm, an alternative approach to document authentication is considered. This approach assigns a unique signature to documents. Because a signature has a numerical value, the database may be organized not as a set of digital images but as a set of signatures according to their numerical values. Certainly, a code representation of images must be such that images can still be identified. Image encoding has many advantages, such as smaller database, easier search through the database, and so on, but it is not clear how to form codes so that codes remain the same (or almost the same) when images are submitted to translation, rotation, and fiber missing effects. In this thesis, three promising methods based on the distances between fiber end points in documents are proposed.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Ahuja, N., and Schachter, B., *Pattern Models,* John Wiley and Sons, 1983.

[2] Alvertos, N., Brzakovic, D., and Gonzalez, R., "Camera Geometries for Image Matching in 3-D Machine Vision," *IEEE Transactions on Pattern Analysis and Machine Vision,* vol. 11, no. 9, 1981.

[3] Gleason, S.S., Hunt, M.A., Jatko, B., "Subpixel Measurement of Image Features Based on Paraboloid Surface Fit," SPIE Proc.: Advances in Intelligent Systems, Boston, MA, 1990.

[4] Gonzalez, R., and Wintz, P., *Digital Image Processing,* Addison-Wesley, 1987.

[5] Goshtasby, A., Stockman, G.C., and Page, C.V., "A Region-Based Approach to Digital Image Registration with Subpixel Accuracy," *IEEE Trans. Geosci. and Remote Sensing,* vol. GE-24, no. 3, 1986.

[6] Keren, D., Peleg, S., and Brada, R., "Image Sequence Enhancement Using Sub-Pixel Displacements," Proc. CVPR '88: The Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 1988, Ann Arbor, MI, 1988.

[7] Longuet-Higgins, M., S., "On the Statistical Distribution of the Heights of Sea Waves," *J. Marine Research XI(3),* 1952.

[8] Longuet-Higgins, M., S., "The Statistical Analysis of a Random Moving Surface," *Phil. Trans. Roy. Soc. London A 249,* 1957.

[9] Longuet-Higgins, M., S., "The Statistical Distribution of the Curvature of a Random Gaussian Surface," *Proc. Cambridge Phil. Soc. 54(4),* 1958.

[10] Longuet-Higgins, M., S., "Statistical Properties of an Isotropic Random Surface," *Phil. Trans. Roy. Soc. London A 250,* 1957.

[11] Oppenheim, A., and Schafer, R., *Discrete-Time Signal Processing*, Prentice Hall, 1989.

[12] Papoulis, A., *Probability, Random Variables and Stochastic Processes*, McGraw Hill, 1984.

[13] Qi, T., and Huhns, M.N., "Algorithms for Sub-Pixel Registration," *Computer Vision, Graphics and Image Processing,* vol. 35, no. 2, 1986.

[14] Sutton., M.A., Mingqui, C., Peters, W.H., Chao, and Y.J., McNeill, S.R., "Application of an Optimized Digital Correlation Method to Planar Deformation Analysis," *Image and Vision Comput.,* vol. 4, no. 3, 1986.

[15] Tobin, K.W., Hunt, M.A., and Jatko, B., "High Resolution Three-Dimensional Image Registration," SPIE Proc.: Applications of Digital Image Processing XII, San Diego, CA, 1989.

Vita

Nenad Vujovic was born in Gornji Milanovac, Yugoslavia on January 24, 1966. He attended elementary schools in that city and was graduated from Victor Nesovic High School in June 1984. In September 1985 he entered Belgrade University, and in January 1991 he received a Bachelor of Science degree in Electrical Engineering. In the fall of 1991 he accepted a research assistantship at The University of Tennessee, Knoxville and began study toward a Master's degree. This degree was awarded in July 1992.