Communication and Cybernetics 10

# Digital Pattern Recognition

Second Corrected and Updated Edition

Edited by K. S. Fu

Springer-Verlag Berlin Heidelberg New York

# Digital Pattern Recognition

Edited by K. S. Fu

With Contributions by
T. M. Cover E. Diday K. S. Fu A. Rosenfeld
J. C. Simon T. J. Wagner J. S. Weszka
J. J. Wolf

Second Corrected and Updated Edition

With 59 Figures

Springer-Verlag
Berlin Heidelberg New York 1980

Professor KING SUN FU, PhD
Purdue University, School of Electrical Engineering
West Lafayette, IN 47907, USA

*Series Editor:*

Professor KING SUN FU, PhD
Purdue University, School of Electrical Engineering
West Lafayette, IN 47907, USA

Professor Dr. WOLF DIETER KEIDEL
I. Physiologisches Institut der Universität Erlangen-Nürnberg
D-8520 Erlangen, Fed. Rep. of Germany

Professor Dr. Willem J. M. Levelt
Max-Planck-Institut für Psycholinguistik, Berg en Dalseweg 79
6522 BC Nijmegen, The Netherlands

ISBN 3-540-10207-8 2. Auflage Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-10207-8 2nd edition Springer-Verlag New York Heidelberg Berlin

ISBN 3-540-07511-9 1. Auflage Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-07511-9 1st edition Springer-Verlag New York Heidelberg Berlin

Library of Congress Cataloging in Publication Data. Main entry under title: Digital pattern recognition. (Communication and cybernetics; 10) Bibliography: p. Includes index. 1. Pattern perception. 2. Automatic speech recognition. I. Fu, King Sun, 1930-Q327.D53 1980 001.53'4 80-20377



Monophoto typesetting, offset printing, and book binding: Brühlsche Universitätsdruckerei, Giessen
2153/3130-543210

# Preface to the Second Edition

Since its publication in 1976, the original volume has been warmly received. We have decided to put out this updated paperback edition so that the book can be more accessible to students. This paperback edition is essentially the same as the original hardcover volume except for the addition of a new chapter (Chapter 7) which reviews the recent advances in pattern recognition and image processing. Because of the limitations of length, we can only report the highlights and point the readers to the literature. A few typographical errors in the original edition were corrected.

We are grateful to the National Science Foundation and the Office of Naval Research for supporting the editing of this book as well as the work described in Chapter 4 and a part of Chapter 7.

West Lafayette, Indiana
March 1980

K. S. Fu

# Preface to the First Edition

During the past fifteen years there has been a considerable growth of interest in problems of pattern recognition. Contributions to the blossom of this area have come from many disciplines, including statistics, psychology, linguistics, computer science, biology, taxonomy, switching theory, communication theory, control theory, and operations research. Many different approaches have been proposed and a number of books have been published. Most books published so far deal with the decision-theoretic (or statistical) approach or the syntactic (or linguistic) approach. Since the area of pattern recognition is still far from its maturity, many new research results, both in theory and in applications, are continuously produced. The purpose of this monograph is to provide a concise summary of the major recent developments in pattern recognition.

The five main chapters (Chapter 2-6) in this book can be divided into two parts. The first three chapters concern primarily with basic techniques in pattern recognition. They include statistical techniques, clustering analysis and syntactic techniques. The last two chapters deal with applications; namely, picture recognition, and speech recognition and understanding. Each chapter is written by one or two distinguished experts on that subject. The editor has not attempted to impose upon the contributors to this volume a uniform notation and terminology, since such notation and terminology does not as yet exist in pattern recognition. Indeed, the diversity of the points of view, notation, and terminology in pattern recognition is a reflection of the fact that this area is in a state of rapid growth and changing orientation. We have not included a chapter on feature (and primitive) extraction and selection, primarily because of the fact that there has been very little major progress made during the last two or three years on this subject. There is no doubt that it is the authors of the individual chapters whose contributions made this volume possible. The editor wishes to express heartfelt thanks to the authors for their cooperation in its rapid completion.

West Lafayette, Indiana
June 1975

K. S. Fu

# Contributors

COVER, THOMAS M.
Stanford Electronics Laboratory, Stanford University, Stanford, CA 94305, USA

DIDAY, EDWIN
Université de Paris IX, Avenue de Pologne, F-75775 Paris Cedex 16

FU, KING SUN
Purdue University, School of Electrical Engineering, West Lafayette, IN 47907, USA

ROSENFELD, AZRIEL
Picture Processing Laboratory, Computer Science Center, University of Maryland, College Park, MD 20742, USA

SIMON, JEAN CLAUDE
Université de Paris VI, Institut de Programmation, Tours 55–56, 4 Place Jussieu, F-75230 Paris Cedex 05

WAGNER, TERRY J.
University of Texas, Department of Electrical Engineering, Austin, TX 78712, USA

WESZKA, JOAN S.
Picture Processing Laboratory, Computer Science Center, University of Maryland, College Park, MD 20742, USA

WOLF, JARED J.
Bolt Beranek and Newman Inc., 50 Moulton Street, Cambridge, MA 02138, USA

# Contents

# 1. Introduction

K. S. Fu

With 2 Figures

## 1.1 What is Pattern Recognition?

The problem of pattern recognition usually denotes a discrimination or classification of a set of processes or events. The set of processes or events to be classified could be a set of physical objects or a set of mental states. The number of pattern classes is often determined by the particular application in mind. For example, consider the problem of English character recognition; we should have a problem of 26 classes. On the other hand, if we are interested in discriminating English characters from Russian characters, we have only a two-class problem. In some problems, the exact number of classes may not be known initially; and it may have to be determined from the observations of many representative patterns, In this case, we would like to detect the possibility of having new classes of patterns as we observe more and more patterns. Human beings perform the task of pattern recognition in almost every instant of their working lives. Recently, scientists and engineers started to use machines for pattern recognition.

An intuitively appealing approach for pattern recognition is the approach of "template-matching". In this case, a set of templates or prototypes, one for each pattern class, is stored in the machine. The input pattern (with unknown classification) is compared with the template of each class, and the classification is based on a preselected matching criterion or similarity criterion. In other words, if the input pattern matches the template of $i$th pattern class better than it matches any other templates, then the input is classified as from the $i$th pattern class. Usually, for the simplicity of the machine, the templates are stored in their raw-data form. This approach has been used for some existing printed-character recognizers and bank-check readers. The disadvantage of the template-matching approach is that it is, sometimes, difficult to select a good template from each pattern class, and to define a proper matching criterion. The difficulty is especially remarkable when large variations and distortions are expected in all the patterns belonging to one class. Recently, the use of flexible template matching or "rubber-mask" techniques has been proposed, see [1.20, 21].

## 1.2 Approaches to Pattern Recognition

The many different mathematical techniques used to solve pattern recognition problems may be grouped into two general approaches; namely, the *decision-theoretic* (or *statistical*) *approach* and the *syntactic* (or *linguistic*) *approach*. In the decision-theoretic approach, a set of characteristic measurements, called features,

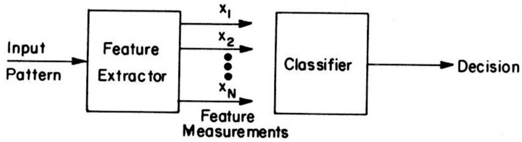


Fig. 1.1. A pattern recognition system

are extracted from the patterns; the recognition of each pattern (assignment to a pattern class) is usually made by partitioning the feature space [1.1]. Most of the developments in pattern recognition research during the past decade deal with the decision-theoretic approach [1.1–18]. Applications include character recognition, crop classification, medical diagnosis, classification of electrocardiograms, etc.

In the decision-theoretic approach, instead of simply matching the input pattern with the templates, the classification is based on a set of selected measurements, extracted from the input pattern. These selected measurements, called "features", are supposed to be invariant or less sensitive with respect to the commonly encountered variations and distortions, and also containing less redundancies. Under this proposition, pattern recognition can be considered as consisting of two subproblems. The first subproblem is what measurements should be taken from the input patterns. Usually, the decision of what to measure is rather subjective and also dependent on the practical situations (for example, the availability of measurements, the cost of measurements, etc.). Unfortunately, at present, there is very little general theory for the selection of feature measurements. However, there are some investigations concerned with the selection of a subset and the ordering of features in a given set of measurements. The criterion of feature selection or ordering is often based on either the importance of the features in characterizing the patterns or the contribution of the features to the performance of recognition (i.e., the accuracy of recognition).

The second subproblem in pattern recognition is the problem of classification (or making a decision on the class assignment to the input patterns) based on the measurements taken from the selected features. The device or machine which extracts the feature measurements from input patterns is called a feature extractor. The device or machine which performs the function of classification is called a classifier. A simplified block diagram of a pattern recognition system is shown in Fig. 1.1 [1]. Thus, in general terms, the template-matching approach may be interpreted as a special case of the second approach—"feature-extraction" approach where the templates are stored in terms of feature measurements and a special classification criterion (matching) is used for the classifier.

In some pattern recognition problems, the structural information which describes each pattern is important, and the recognition process includes not only the capability of assigning the pattern to a particular class (to classify it), but also the capacity to describe aspects of the pattern which make it ineligible for assignment to another class. A typical example of this class of recognition problem is picture recognition or more generally speaking, scene analysis. In this class of recognition problems, the patterns under consideration are usually quite complex

[1] The division into two parts is primarily for convenience rather than necessity.

and the number of features required is often very large which makes the idea of describing a complex pattern in terms of a (hierarchical) composition of simpler subpatterns very attractive. Also, when the patterns are complex and the number of possible descriptions is very large, it is impractical to regard each description as defining a class (for example, in fingerprint and face identification problems, recognition of continuous speech, Chinese characters, etc.). Consequently, the requirement of recognition can only be satisfied by a description for each pattern rather than the simple task of classification. In order to represent the hierarchical (tree-like) structural information of each pattern, that is, a pattern described in terms of simpler subpatterns and each simpler subpattern again be described in terms of even simpler subpatterns, etc., the *syntactic* or *structural approach* has been proposed [1.19].

After a very brief summary of some basic pattern recognition methods in this chapter[2] a review of recent progress in both the decision-theoretic approach and the syntactic approach is presented. In the decision-theoretic approach, statistical methods and cluster analysis are discussed. In addition, picture recognition and speech recognition and understanding are included in this volume to demonstrate general applications of pattern recognition methods to the processing of one-dimensional speech signals and two-dimensional pictorial patterns.

## 1.3 Basic Non-Parametric Decision—Theoretic Classification Methods

The concept of pattern classification may be expressed in terms of the partition of feature space (or a mapping from feature space to decision space). Suppose that $N$ features are to be measured from each input pattern. Each set of $N$ features can be considered as a vector $X$, called a feature (measurement) vector, or a point in the $N$-dimensional feature space $\Omega_X$. The problem of classification is to assign each possible vector or point in the feature space to a proper pattern class. This can be interpreted as a partition of the feature space into mutually exclusive regions, and each region will correspond to a particular pattern class. Mathematically, the problem of classification can be formulated in terms of "discriminant functions". Let $\omega_1, \omega_2, \ldots, \omega_m$ be designated as the $m$ possible pattern classes to be recognized, and let

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \tag{1.1}$$

be the feature (measurement) vector where $x_i$ represents the $i$th feature measurement. Then the discriminant function $D_j(X)$ associated with pattern class $\omega_j$,

[2] For more extensive discussions of existing pattern recognition methods and their applications, refer to [1.1–19 and 1.23–27].

$j=1,\ldots,m$, is such that if the input pattern represented by the feature vector $X$ is in class $\omega_i$, denoted as $X \sim \omega_i$, the value of $D_i(X)$ must be the largest. That is, for all $X \sim \omega_i$,

$$D_i(X) > D_j(X), \quad i,j=1,\ldots,m, \quad i \neq j. \tag{1.2}$$

Thus, in the feature space $\Omega_X$ the boundary of partition, called the decision boundary, between regions associated with class $\omega_i$ and class $\omega_j$, respectively, is expressed by the following equation.

$$D_i(X) - D_j(X) = 0. \tag{1.3}$$

Many different forms satisfying condition (1.2) can be selected for $D_i(X)$. Several important discriminant functions are discussed in the following.

### 1.3.1 Linear Discriminant Functions

In this case, a linear combination of the feature measurements $x_1, x_2, \ldots, x_N$ is selected for $D_i(X)$, i.e.,

$$D_i(X) = \textstyle\sum_{k=1}^{N} w_{ik} x_k + w_{i,N+1}, \quad i=1,\ldots,m. \tag{1.4}$$

The decision boundary between regions in $\Omega_X$ associated with $\omega_i$ and $\omega_j$ is in the form of

$$D_i(X) - D_j(X) = \textstyle\sum_{k=1}^{N} w_k x_k + w_{N+1} = 0 \tag{1.5}$$

with $w_k = w_{ik} - w_{jk}$ and $w_{N+1} = w_{i,N+1} - w_{j,N+1}$. Equation (1.5) is the equation of a hyperplane in the feature space $\Omega_X$. If $m=2$, on the basis of (1.5), $i,j=1,2$ $(i \neq j)$, a threshold logic device, as shown in Fig. 1.2, can be employed as a linear classifier (a classifier using linear discriminant functions). From Fig. 1.2, let $D(X) = D_1(X) - D_2(X)$, if

$$\begin{aligned} &\text{output} = +1, \quad \text{i.e.,} \quad D(X) > 0, \quad X \sim \omega_1 \\ \text{and if } \; &\text{output} = -1, \quad \text{i.e.,} \quad D(X) < 0, \quad X \sim \omega_2 . \end{aligned} \tag{1.6}$$

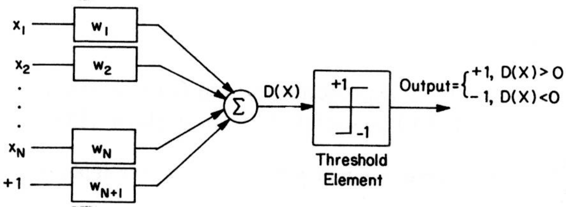


Fig. 1.2. A linear two-class classifier

For the number of pattern classes more than two, $m>2$, several threshold logic devices can be connected in parallel so that the combinations of the outputs from, say, $M$ threshold logic devices will be sufficient for distinguishing $m$ classes, i.e., $2^M \geqq m$.

### 1.3.2 Minimum Distance Classifier

An important class of linear classifiers is that of using the distances between the input pattern and a set of reference vectors or prototype points in the feature space as the classification criterion. Suppose that $m$ reference vectors $R_1, R_2, \ldots, R_m$, are given with $R_j$ associated with the pattern class $\omega_j$. A minimum-distance classification scheme with respect to $R_1, R_2, \ldots, R_m$ is to classify the input $X$ as from class $\omega_i$, i.e.,

$$X \sim \omega_i \quad \text{if} \quad |X - R_i| \text{ is the minimum}, \tag{1.7}$$

where $|X - R_i|$ is the distance defined between $X$ and $R_i$. For example, $|X - R_i|$ may be defined as

$$|X - R_i| = \sqrt{(X - R_i)^T (X - R_i)}, \tag{1.8}$$

where the superscript $T$ represents the transpose operation to a vector. From (1.8).

$$|X - R_i|^2 = X^T X - X^T R_i - X R_i^T + R_i^T R_i . \tag{1.9}$$

Since $X^T X$ is not a function of $i$, the corresponding discriminant function for a minimum-distance classifier is essentially

$$D_i(X) = X^T R_i + X R_i^T - R_i^T R_i, \qquad i = 1, \ldots, m \tag{1.10}$$

which is linear. Hence, a minimum-distance classifier is also a linear classifier. The performance of a minimum-distance classifier is, of course, dependent upon an appropriately selected set of reference vectors.

### 1.3.3 Piecewise Linear Discriminant Functions (Nearest Neighbor Classification)

The concept adopted in Subsection 1.3.2 can be extended to the case of minimum-distance classification with respect to sets of reference vectors. Let $R_1, R_2, \ldots, R_m$ be the $m$ sets of reference vectors associated with classes $\omega_1, \omega_2, \ldots, \omega_m$, respectively, and let reference vectors in $R_j$ be denoted as $R_j^{(k)}$, i.e.,

$$R_j^{(k)} \in R_j, \qquad k = l, \ldots, u_j,$$

where $u_j$ is the number of reference vectors in set $R_j$. Define the distance between an input feature vector $X$ and $R_j$ as

$$d(X, R_j) = \min_{k=1,\ldots,u_j} |X - R_j^{(k)}| . \tag{1.11}$$

That is, the distance between $X$ and $R_j$ is the smallest of the distances between $X$ and each vector in $R_j$. The classifier will assign the input to a pattern class which is associated with the closest vector set. If the distance between $X$ and $R_j^{(k)}$, $|X-R_j^{(k)}|$ is defined as (1.8), then the discriminant function used in this case is essentially

$$D_i(X) = \max_{k=1,\ldots,u_i} \{X^{\mathrm{T}}R_i^{(k)} + (R_i^{(k)})^{\mathrm{T}}X - (R_i^{(k)})^{\mathrm{T}}R_i^{(k)}\}, \qquad i=1,\ldots,m. \tag{1.12}$$

Let

$$D_i^{(k)} = X^{\mathrm{T}}R_i^{(k)} + (R_i^{(k)})^{\mathrm{T}}X - (R_i^{(k)})^{\mathrm{T}}R_i^{(k)}. \tag{1.13}$$

Then

$$D_i(X) = \max_{k=1,\ldots,u_i} \{D_i^{(k)}(X)\}, \qquad i=1,\ldots,m. \tag{1.14}$$

It is noted that $D_i^{(k)}(X)$ is a linear combination of features, hence, the class of classifiers using (1.12) or (1.14) is often called piecewise linear classifiers.

### 1.3.4 Polynomial Discriminant Functions

An $r$th-order polynomial discriminant function can be expressed as

$$D_i(X) = w_{i1}f_1(X) + w_{i2}f_2(X) + \ldots + w_{iL}f_L(X) + w_{i,L+1} \tag{1.15}$$

where $f_j(X)$ is of the form

$$x_{k_1}^{n_1}x_{k_2}^{n_2}\ldots x_{k_r}^{n_r} \quad \text{for} \quad \begin{array}{l} k_1, k_2, \ldots, k_r = 1, \ldots, N, \quad \text{and} \\ n_1, n_2, \ldots, n_r = 0 \text{ and } 1. \end{array} \tag{1.16}$$

The decision boundary between any two classes is also in the form of an $r$th-order polynomial. Particularly, if $r=2$, the discriminant function is called a quadric discriminant function. In this case,

$$f_j(X) = x_{k_1}^{n_1}x_{k_2}^{n_2} \quad \text{for} \quad k_1, k_2 = 1, \ldots, N, \quad \text{and} \quad n_1, n_2 = 0 \text{ and } 1 \tag{1.17}$$

and

$$L = \tfrac{1}{2}N(N+3). \; = \tfrac{1}{2}(N \times N - N) + N \tag{1.18}$$

Typically,

$$D_i(X) = \textstyle\sum_{k=1}^{N} w_{kk}x_k^2 + \sum_{j=1}^{N-1}\sum_{k=j+1}^{N} w_{jk}x_jx_k + \sum_{j=1}^{N} w_jx_j + w_{L+1}. \tag{1.19}$$

In general, the decision boundary for quadric discriminant functions is a hyper-hyperboloid. Special cases include hypersphere, hyperellipsoid and hyperellipsoidal cylinder.