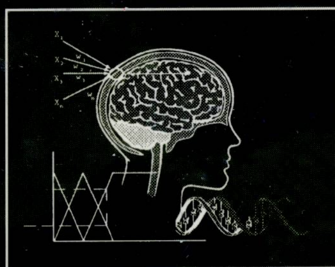
Edited by

Lakhmi C. Jain

Beatrice Lazzerini

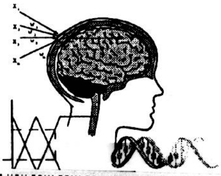


KNOWLEDGE-BASED INTELLIGENT TECHNIQUES in CHARACTER RECOGNITION

KNOWLEDGE-BASED INTELLIGENT TECHNIQUES in CHARACTER RECOGNITION

Edited by

Lakhmi C. Jain
Beatrice Lazzerini



CRC Press
Boca Raton London New York Washington, D.C.

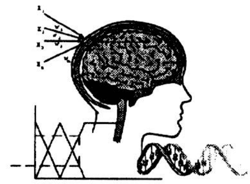
Library of Congress Cataloging-in-Publication Data

Knowledge-based intelligent techniques in character recognition /
[edited by] L.C. Jain, B. Lazzerini.
p. cm. -- (The CRC Press international series on
computational intelligence)
Includes bibliographical references and index.
ISBN 0-8493-9807-X (alk. paper)
1. Optical character recognition devices. 2. Writing-
-Identification--Data processing. 3. Expert systems. 4. Soft
computing. I. Jain, L. C. II. Lazzerini, Beatrice, 1953– .
III. Series.
TA1640.K56 1999
006.4'24—dc21 99-12428
CIP

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

International Standard Book Number 0-8493-9807-X
Library of Congress Card Number 99-12428
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper



The CRC Press

International Series on Computational Intelligence

Series Editor
L.C. Jain, Ph.D., M.E., B.E., (Hons), Fellow I.E. (Australia)

L.C. Jain, R.P. Johnson, Y. Takefuji, and L.A. Zadeh
Knowledge-Based Intelligent Techniques in Industry

L.C. Jain and C.W. de Silva
Intelligent Adaptive Control: Industrial Applications

L.C. Jain and N.M. Martin
Fusion of Neural Networks, Fuzzy Systems, and Genetic Algorithms: Industrial Applications

H.N. Teodorescu, A. Kandel, and L.C. Jain
Fuzzy and Neuro-Fuzzy Systems in Medicine

C.L. Karr and L.M. Freeman
Industrial Applications of Genetic Algorithms

L.C. Jain and Beatrice Lazzerini
Knowledge-Based Intelligent Techniques in Character Recognition

L.C. Jain and V. Vemuri
Industrial Applications of Neural Networks

H.N. Teodorescu, A. Kandel, and L.C. Jain
Soft Computing in Human-Related Science

B. Lazzerini, D. Dumitrescu, L.C. Jain, and A. Dumitrescu
Evolutionary Computing and Applications

B. Lazzerini, D. Dumitrescu, and L.C. Jain
Fuzzy Sets and Their Application to Clustering and Training

L.C. Jain, U. Halici, I. Hayashi, S.B. Lee, and S. Tsutsui
Intelligent Biometric Techniques in Fingerprint and Face Recognition: Practical Applications

Z. Chen
Computational Intelligence for Decision Support

L.C. Jain
Evolution of Engineering and Information System and Their Applications

H.N. Teodorescu and A. Kandel
Dynamic Fuzzy Systems and Chaos Applications

L. Medsker and L.C. Jain
Recurrent Neural Networks: Design and Applications

PREFACE

The main aim of this book is to present results of research on intelligent character recognition techniques. There is tremendous worldwide interest in the development and application of personal identification techniques. Personal identification techniques are useful in many situations including security, bank check processing and postal code recognition. Examples of personal identification techniques are character recognition, fingerprint recognition, face recognition and iris recognition. Most character recognition techniques involve data acquisition, preprocessing, feature extraction and classification.

The main objective of automating the character recognition process has been to make the automated system more 'human like.' The primary techniques employed to achieve 'human like' behavior have been fuzzy logic, neural networks and evolutionary computing.

This book has eleven chapters. The first chapter, by Jain and Lazzerini, introduces principles of signature verification for on-line and off-line systems. Some considerations regarding performance evaluation of handwriting recognition systems are presented. The second chapter, by Shouno, Fukushima and Okada, shows that neocognitrons trained by unsupervised learning can recognize real-world handwritten digits provided by a large database. The third chapter, by Satoh, Kuroiwa, Aso and Miyake, presents a novel rotation-invariant neocognitron. This rotation-invariant recognition is implemented by extending the neocognitron, which can recognize translated, scaled and/or distorted patterns from those used in training.

The fourth chapter, by Baldwin, Martin and Stylianidis, discusses the integration of soft computing to develop a system for recognizing hand-written numerals. The fifth chapter, by Sorbello, Gioiello and Vitabile, presents algorithms for recognizing handwritten characters using a multilayer perceptron neural network. The sixth chapter, by Liu and Fung, presents a technique for automatically identifying signature features and verifying signatures with greater certainty. The use of fuzzy-genetic algorithms overcomes traditional problems in feature classification and selection, providing fuzzy templates for the identification of the smallest subset of features.

The seventh chapter, by Banarse and Duller, presents the application of a self-organizing neural network to handwritten character recognition. The eighth chapter, by Lazzerini, Reyneri, Gregoretti and Mariani, presents an Italian bank check processing system. This system consists of several processing modules including those for data acquisition, image preprocessing, character center detection, character recognition, courtesy amount recognition and legal amount recognition. The ninth chapter, by El-Yacoubi, Sabourin, Gilloux and Suen, presents an off-line handwritten word recognition system using hidden Markov models. Chapter ten, by Malaviya, Ivancic, Balasubramaniam and Peters, presents off-line handwriting recognition systems using context-dependent fuzzy rules. The last chapter, by Brugge, Nijhuis, Spaanenburg and Stevens, presents a license-plate recognition system for automated traffic monitoring and law enforcement on public roads.

This book will be useful to researchers, practicing engineers and students who wish to develop a successful character recognition system.

We would like to express our sincere thanks to Berend Jan van der Zwaag, Ashlesha Jain, Ajita Jain and Sandhya Jain for their assistance in the preparation of the manuscript. We are grateful to the authors for their contributions, and thanks are due to Dawn Mesa, Mimi Williams, Lourdes Franco and Suzanne Lassandro for their editorial assistance.

L.C. Jain, Australia
B. Lazzerini, Italy

CONTENTS

Chapter 1:

An Introduction to Handwritten Character and Word Recognition

AN INTRODUCTION TO HANDWRITTEN CHARACTER AND WORD RECOGNITION

L.C. Jain
Knowledge-Based Intelligent Engineering Systems Centre
University of South Australia
Adelaide, Mawson Lakes, SA 5095, Australia

B. Lazzerini
Dipartimento di Ingegneria della Informazione
Università degli Studi di Pisa
Via Diotisalvi 2, 56126 Pisa, Italy

There is a worldwide interest in the development of handwritten character and word recognition systems. These systems are used in many situations such as recognition of postcodes, interpretation of amount and verification of signature on bank checks, and law enforcement on public roads. The tremendous advances in the computational intelligence techniques have provided new tools for the development of intelligent character recognition systems. This chapter introduces principles of handwriting recognition for on-line and off-line systems. Some considerations about performance evaluation of a handwriting recognition system are discussed.

1 Introduction

In the last few years many academic institutions and industrial companies have been involved in the field of handwriting recognition. The automatic recognition of handwritten text can be extremely useful in many applications where it is necessary to process large volumes of

0-8493-9807-X/99/$0.00+$.50

handwritten data, such as recognition of addresses and postcodes on envelopes, interpretation of amounts on bank checks, document analysis, and verification of signatures [1]. Substantial progress has been recently achieved, but the recognition of handwritten text cannot yet approach human performance. The major difficulties descend from the variability of someone's calligraphy over time, the similarity of some characters with each other, and the infinite variety of character shapes and writing styles produced by different writers. Furthermore, the possible low quality of the text image, the unavoidable presence of background noise and various kinds of distortions (such as poorly written, degraded, or overlapping characters) can make the recognition process even more difficult. Therefore, handwriting recognition is still an open and interesting area for research and novel ideas.

Handwriting recognition systems acquire data by means of either off-line devices (such as scanners and cameras) or on-line devices (such as graphic tablets). In off-line recognition systems, the input image is converted into a bit pattern. Then, specific preprocessing algorithms prepare the acquired data for subsequent processing by eliminating noise and errors caused by the acquisition process. In contrast, on-line recognition systems use dynamic writing information, namely, the number, order, and direction of the strokes to recognize each character as it is being written. Typically, the two-dimensional coordinates of the pen position on the writing surface are sampled at fixed time intervals.

The field of automatic recognition of handwritten isolated *characters* is the longest established branch of handwriting recognition. A character is traditionally identified by the set of its fundamental components, or *features*. A great variety of types of features have been used, including topological and geometrical, directional, mathematical, and structural features. Several conventional and computational intelligence techniques have been applied to character recognition such as statistical analysis, structural and syntactic analysis, pattern matching, neural networks, fuzzy systems and genetic algorithms.

A more difficult field is the recognition of handwritten *words*, especially fully cursive (i.e., with a fully connected main body) words. Two approaches to word recognition are mainly used; the global approach and the analytical approach. The *global* (or *holistic*) approach recognizes the whole word. Each word is represented in terms of global features (such as the sequence and position of vertical strokes, ascenders, descenders, and loops) and the representation of the word is compared with models, describing the ideal shape of the possible words, stored in a reference lexicon [2]. One or more models may be built for each word to take into account the wide variability of cursive script. The global approach is usually adopted in applications with a reduced size (not exceeding a few hundred words) lexicon (for example, in check processing).

The *analytical* approach divides a word into smaller units, typically characters or groups of characters (*segmentation* process). The isolated units are then individually recognized and fit within a possible word.

Of course, the recognition accuracy of a handwriting recognition system is strongly dependent on the segmentation correctness; particularly, an erroneous segmentation may lead to incorrect recognition. A wrong segmentation may be produced, for example, by the lack of a clear separation between consecutive characters, or by the fact that the same sequence of strokes can be interpreted (also by human readers) in different ways.

To solve, at least partially, the aforementioned problems, a third approach to word recognition, named *integrated segmentation and recognition* (or *recognition-based* approach), is sometimes adopted. It consists in tightly integrating word segmentation with character recognition, without requiring any preliminary word segmentation.

It is well known that robust and efficient handwriting recognition can only be achieved by the use of *context*; i.e., information external to the characters [3], [4], [5]. In fact, even humans can make mistakes when reading without utilizing context because of the geometric diversity of character shapes produced by different writers. The problem is even worse when reading handwritten text by computer owing to imperfect writing instruments, scanning mechanisms, learning techniques and recognition algorithms.

Using context means exploiting domain-specific information to make the recognition process easier. There are basically three types of knowledge: morphological, pragmatic and linguistic [6]. *Morphological* knowledge refers to the shape of an ideal (i.e., writer-independent) representation of the characters. *Pragmatic* knowledge refers to the spatial arrangement of ideal character representations into words and phrases. *Linguistic* knowledge refers to the specific language used and concerns the lexical, grammatical, and semantic aspects. In particular, syntax and semantics are used to describe how words and phrases in a document are related. Pragmatic and linguistic knowledge can be considered contextual knowledge.

Examples of the use of contextual knowledge are the following. Simple reading mistakes can be detected and corrected by verifying recognized words through a dictionary, the recognition of handwritten ZIP codes can be made easier by verifying the digits with the address information, etc. Contextual knowledge can also be used to ensure proper word identification in the case in which only parts of the handwritten text image have been recognized because of missing characters, noise or distortions.

2 Knowledge-Based Intelligent Techniques

The knowledge-based intelligent techniques [7] include expert systems, neural networks, fuzzy systems and evolutionary computing. These techniques try to mimic the performance of biological systems in a very limited sense. Expert systems solve specific problems in a given domain in the same way as human experts. Artificial neural networks are designed to mimic the biological information processing mechanism. Evolutionary computing is used in implementing many functions of intelligent systems including optimisation. Fuzzy logic provides a basis for representing uncertain and imprecise knowledge. The majority of chapters in this book demonstrate the application of knowledge-based intelligent techniques in character recognition.

3 Off-Line Handwriting Recognition

Most current approaches to off-line handwriting recognition consist of three main phases, namely, *preprocessing*, *segmentation*, and *recognition*. In describing these phases, we will refer to a generic passage of text. Furthermore, being convinced that an automatic reading system should take soft decisions (as the human cortex does), we assume that most modules (like the character and word recognizers) of such a system would output a list of candidates, possibly ordered by decreasing confidence values, instead of just one.

3.1 Preprocessing

The preprocessing phase aims to extract the relevant textual parts and prepares them for segmentation and recognition. (Note that, depending on the adopted method, some of the preprocessing actions described below may be omitted as well as performed in a slightly different order). First of all, the image may be physically rotated and/or shifted for slant and skew correction. To this aim, the average slope of horizontal/vertical strokes in the specific text is estimated using a Hough transformation of the contour or by computing directional histograms of pixel densities. Then the reference base line of each text line is found, which detects the area where the handwritten text is located.

To remove noise from the image, the document image is filtered. Then, it is thresholded to leave every pixel black or white. A thinning algorithm is finally used to reduce the width of lines to 1 pixel while preserving stroke connectivity. Thinning finds the skeleton of a character image to make identification of character features easier. The raw skeletons are then smoothed by removing edge points.

The next crucial point is the removal of printed boxes and horizontal guidelines used to indicate the position of the information fields to be filled in manually by the writer. Examples of such fields are the courtesy (digit) amount, the legal (worded) amount and the customer's signature fields in bank checks. As the guidelines can be overlapped by the handwritten text, some textual pixels could be deleted as well. A connected stroke or line segment might be broken into two or several parts, resulting, for instance, in the letter "b" to be transformed into "h". This kind of ambiguity is solved by matching with words in a dictionary.

Several different approaches to guideline removal have been proposed that limit handwritten text erosion [8], [9]. Typical examples are the morphological erosion and dilation operators [10]. In general, however, simple heuristic criteria can help to distinguish guidelines from text. For example, variations in the average direction of the guideline indicate the presence of pixels belonging to the text.

Horizontal density histograms can also be used to detect the upper, middle, and lower zones of a word. The first and last zones contain, respectively, ascenders and descenders. The second zone contains characters like "a", "c", "e", etc. Word zoning information is useful to determine the type of a word, that is, whether it is made of middle zone letters only or it also includes letters with ascenders and/or descenders. Also, singularities of handwriting (corresponding to key letters) can be detected by using vertical histograms within the upper or lower zone. Similar techniques can be used to estimate the character height to provide scale invariance. Finally, conversion techniques from two-dimensional to one-dimensional data could be adopted to exploit on-line recognition methods.

3.2 Segmentation

The text lines are segmented into sentences, words and characters. Sentences are supposed to be terminated by punctuation marks. Typically, potential cuts into words of a given sentence are generated based on the distances between adjacent characters (of course, a blank space is a clue to word separation). The probability (or confidence) for each potential cut to be a correct sentence segmentation is calculated. Then, one or more possible sentence segmentation options are considered. Finally, the most probable segmentation options give origin to sequences of possible words. Each such word is input to the word recognizer. The word recognizer, exploiting one of the techniques previously introduced, outputs an ordered list of the best n words (candidate words) in the lexicon which correspond to the unknown word. The list is sorted according to word confidence values. Of course, the word recognizer may contain a segmentation process at the character level.

For each sentence segmentation option, a list of candidate sentences is generated from the combination of the corresponding candidate words and verified by means of syntactic/contextual analysis. Sentence confidence values are associated with the resulting correct sentences.

Of course, the confidence value associated with an entity (sentence, word, etc.) should take into account the confidence values associated with its constituent sub-units.

3.3 Recognition

Essentially, this phase aims to recognize characters and words (note that most applications of handwriting recognition involve text consisting of just one or a few words). In what follows, we will be concerned with word recognition (Sections 3.3.1, 3.3.2, 3.3.3) and character recognition (Section 3.3.4). As previously stated, a word recognizer can apply one of three possible strategies, namely, analytical, holistic and recognition-based [11], [12].

3.3.1 Analytical Methods

The analytical approach segments a word into smaller, easier to recognize units. These units can be of different types according to the adopted method: segments, characters, graphemes, pseudo-characters, etc. It is essential that these units be strictly related to characters so that recognition does not depend on a specific vocabulary.

The analytical approach performs an *explicit* (or *external*) segmentation, because words are explicitly divided into characters (or other units) which are recognized separately. Of course, context-based post-processing is usually required to ensure proper word recognition. Potential breaking points (like ligatures between characters) are found, mainly based on heuristic rules derived from visual observation. Ligatures may be detected using techniques that identify the difference