CHAPTER 1

INTRODUCTION: CHARACTER RECOGNITION, EVOLUTION, AND DEVELOPMENT

This chapter presents an overview of the problems associated with character recognition. It includes a brief description of the history of OCR (optical character recognition), the extensive efforts involved to make it work, the recent international activities that have stimulated the growth of research in handwriting recognition and document analysis, a short summary of the topics to be discussed in the other chapters, and a list of relevant references.

1.1 GENERATION AND RECOGNITION OF CHARACTERS

Most people learn to read and write during their first few years of education. By the time they have grown out of childhood, they have already acquired very good reading and writing skills, including the ability to read most texts, whether they are printed in different fonts and styles, or handwritten neatly or sloppily. Most people have no problem in reading the following: light prints or heavy prints; upside down prints; advertisements in fancy font styles; characters with flowery ornaments and missing parts; and even characters with funny decorations, stray marks, broken, or fragmented parts; misspelled words; and artistic and figurative designs. At times, the characters and words may appear rather distorted and yet, by experience and by context, most people can still figure them out. On the contrary, despite more than five decades of intensive research, the reading skill of the computer is still way behind that of human

Character Recognition Systems: A Guide for Students and Practitioner, by M. Cheriet, N. Kharma, C.-L. Liu and C. Y. Suen

beings. Most OCR systems still cannot read degraded documents and handwritten characters/words.

1.2 HISTORY OF OCR

To understand the phenomena described in the above section, we have to look at the history of OCR [3, 4, 6], its development, recognition methods, computer technologies, and the differences between humans and machines [1, 2, 5, 7, 8].

It is always fascinating to be able to find ways of enabling a computer to mimic human functions, like the ability to read, to write, to see things, and so on. OCR research and development can be traced back to the early 1950s, when scientists tried to capture the images of characters and texts, first by mechanical and optical means of rotating disks and photomultiplier, flying spot scanner with a cathode ray tube lens, followed by photocells and arrays of them. At first, the scanning operation was slow and one line of characters could be digitized at a time by moving the scanner or the paper medium. Subsequently, the inventions of drum and flatbed scanners arrived, which extended scanning to the full page. Then, advances in digital-integrated circuits brought photoarrays with higher density, faster transports for documents, and higher speed in scanning and digital conversions. These important improvements greatly accelerated the speed of character recognition and reduced the cost, and opened up the possibilities of processing a great variety of forms and documents. Throughout the 1960s and 1970s, new OCR applications sprang up in retail businesses, banks, hospitals, post offices; insurance, railroad, and aircraft companies; newspaper publishers, and many other industries [3, 4].

In parallel with these advances in hardware development, intensive research on character recognition was taking place in the research laboratories of both academic and industrial sectors [6, 7]. Although both recognition techniques and computers were not that powerful in the early days (1960s), OCR machines tended to make lots of errors when the print quality was poor, caused either by wide variations in typefonts and roughness of the surface of the paper or by the cotton ribbons of the typewriters [5]. To make OCR work efficiently and economically, there was a big push from OCR manufacturers and suppliers toward the standardization of print fonts, paper, and ink qualities for OCR applications. New fonts such as OCRA and OCRB were designed in the 1970s by the American National Standards Institute (ANSI) and the European Computer Manufacturers Association (ECMA), respectively. These special fonts were quickly adopted by the International Standards Organization (ISO) to facilitate the recognition process [3, 4, 6, 7]. As a result, very high recognition rates became achievable at high speed and at reasonable costs. Such accomplishments also brought better printing qualities of data and paper for practical applications. Actually, they completely revolutionalized the data input industry [6] and eliminated the jobs of thousands of keypunch operators who were doing the really mundane work of keying data into the computer.

1.3 DEVELOPMENT OF NEW TECHNIQUES

As OCR research and development advanced, demands on handwriting recognition also increased because a lot of data (such as addresses written on envelopes; amounts written on checks; names, addresses, identity numbers, and dollar values written on invoices and forms) were written by hand and they had to be entered into the computer for processing. But early OCR techniques were based mostly on template matching, simple line and geometric features, stroke detection, and the extraction of their derivatives. Such techniques were not sophisticated enough for practical recognition of data handwritten on forms or documents. To cope with this, the Standards Committees in the United States, Canada, Japan, and some countries in Europe designed some handprint models in the 1970s and 1980s for people to write them in boxes [7]. Hence, characters written in such specified shapes did not vary too much in styles, and they could be recognized more easily by OCR machines, especially when the data were entered by controlled groups of people, for example, employees of the same company were asked to write their data like the advocated models. Sometimes writers were asked to follow certain additional instructions to enhance the quality of their samples, for example, write big, close the loops, use simple shapes, do not link characters, and so on. With such constraints, OCR recognition of handprints was able to flourish for a number of years.

1.4 RECENT TRENDS AND MOVEMENTS

As the years of intensive research and development went by, and with the birth of several new conferences and workshops such as IWFHR (International Workshop on Frontiers in Handwriting Recognition),[1] ICDAR (International Conference on Document Analysis and Recognition),[2] and others [8], recognition techniques advanced rapidly. Moreover, computers became much more powerful than before. People could write the way they normally did, and characters need not have to be written like specified models, and the subject of unconstrained handwriting recognition gained considerable momentum and grew quickly. As of now, many new algorithms and techniques in preprocessing, feature extraction, and powerful classification methods have been developed [8, 9]. Further details can be found in the following chapters.

1.5 ORGANIZATION OF THE REMAINING CHAPTERS

Nowadays, in OCR, once a printed or handwritten text has been captured optically by a scanner or some other optical means, the digital image goes through the following stages of a computer recognition system:

[1] Note that IWFHR has been promoted to an international conference, namely, the International Conference on Frontiers on Handwriting Recognition (ICFHR), starting in 2008 in Montreal, where it was born in 1990.

[2] IWFHR and ICDAR series were founded by Dr. Ching Y. Suen, coauthor of this book, in 1990 and 1991, respectively.

1. The preprocessing stage that enhances the quality of the input image and locates the data of interest.
2. The feature extraction stage that captures the distinctive characteristics of the digitized characters for recognition.
3. The classification stage that processes the feature vectors to identify the characters and words.

Hence, this book is organized according to the above sequences.

REFERENCES

1. H. Bunke and P. S. P. Wang. *Handbook of Character Recognition and Document Image Analysis*. World Scientific Publishing, Singapore, 1997.
2. S. Mori, H. Nishida, and H. Yamada. *Optical Character Recognition*, Wiley Interscience, New Jersey, 1999.
3. *Optical Character Recognition and the Years Ahead.* The Business Press, Elmhurst, IL, 1969.
4. Pas d'auteur. *Auerbach on Optical Character Recognition*. Auerbach Publishers, Inc., Princeton, 1971.
5. S. V. Rice, G. Nagy, and T. A. Nartker. *Optical Character Recognition: An Illustrated Guide to the Frontier*. Kluwer Academic Publishers, Boston, 1999.
6. H. F. Schantz. *The History of OCR*. Recognition Technologies Users Association, Boston, 1982.
7. C. Y. Suen. Character recognition by computer and applications. In T. Y. Young and K. S. Fu, editors, *Handbook of Pattern Recognition and Image Processing*. Academic Press, Inc., Orlando, FL, 1986, pp. 569–586.
8. Proceedings of the following international workshops and conferences:

   - *ICPR—International Conference on Pattern Recognition*
   - *ICDAR—International Conference on Document Analysis and Recognition*
   - *DAS—Document Analysis Systems*
   - *IWFHR—International Workshop on Frontiers in Handwriting Recognition.*

9. Journals, in particular:

   - *Pattern Recognition*
   - *Pattern Recognition Letters*
   - *Pattern Analysis and Applications*
   - *International Journal on Document Analysis and Recognition*
   - *International Journal of Pattern Recognition and Artificial Intelligence.*