Pattern Recognition 41 (2008) 2435 – 2446

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Forty years of research in character and document recognition—an industrial perspective

Hiromichi Fujisawa*

Central Research Laboratory, Hitachi, Ltd., 1-280 Higashi-koigakubo, Kokubunji, Tokyo 185-8601, Japan

ARTICLE INFO

Article history:
Received 15 February 2008
Received in revised form 10 March 2008
Accepted 11 March 2008

Keywords:
OCR
Character recognition
Handwriting recognition
Kanji recognition
Postal address recognition
Robustness design
Information integration
Hypothesis-driven approaches
Digital pen

ABSTRACT

This paper presents an overview on the last 40-years of technical advances in the field of character and document recognition. Representative developments in each decade are described. Then, key technical developments in the specific area of Kanji recognition in Japan are highlighted. The main part of the paper discusses robustness design principles, which have proven to be effective to solve complex problems in postal address recognition. Included are the hypothesis-driven principle, deferred decision/multiple-hypotheses principle, information integration principle, alternative solution principle, and perturbation principle. Finally, future prospects, the 'long-tail' phenomena, and promising new applications are discussed.

1. Introduction

Presented is an industrial view on the character and document recognition technology, based on some material presented at ICDAR [1]. Commercial optical character readers (OCRs) emerged in the 1950s, and since then, the character and document recognition technology has advanced significantly providing products and systems to meet industrial and commercial needs throughout the development process. At the same time, the profits from businesses based on this technology have been invested in research and development of more advanced technology. We can observe here a virtuous cycle. New technologies have enabled new applications, and the new applications have supported the development of better technology. Character and document recognition has been a very successful area of pattern recognition.

The main business and industrial applications of character and document recognition in the last forty years have been in form reading, bank check reading and postal address reading. By supporting these applications, recognition capability has expanded in multiple dimensions: mode of writing, scripts, types of documents, and so on. The recognizable modes of writing are machine-printing, handprinting, and script handwriting. Recognizable scripts started with Arabic numerals and expanded to the Latin alphabets, Japanese Katakana syllabic characters, Kanji (Japanese version of Chinese) characters, Chinese characters, and Hangul characters. Work is now being done to make Indian and Arabic scripts readable. Many different kinds of paper forms can be read by today's OCRs, including bank checks, post cards, envelopes, book pages, and business cards. Typeface standards such as OCR-A and OCR-B fonts have contributed to making OCRs reliable enough even in the early stages. In the same context, specially designed OCR forms have simplified the segmentation problem and made handprinted character OCRs readable even by immature recognition technology. Today's OCRs are successfully used to read any type of fonts and freely handwritten characters.

The field of character and document recognition has not always been peaceful. It has twice been disturbed by waves of new digital technologies that threatened to diminish the role of OCR technology. The first such wave was that of office automation in the early 1980s. Starting then, most of information seemed to be going to be 'born digital', potentially diminishing demand for OCRs, and some researchers were pessimistic about the future. However, it turned out that the sales of OCRs in Japan, for example, peaked in the 1980s. This was ironically due to the promoted introduction of office computers. It is well known that the use of paper has kept increasing.

We are now facing the second wave. IT and Web technologies might have a different impact. Many kinds of applications can now be completed on the Web. Information can flow around the world in an instant. However, it is still not known whether the demand for character and document recognition will decrease or whether new applications requiring more advanced technology will be created. Search

* Tel.: +81 42 323 1111; fax: +81 42 327 7700.
E-mail address: hiromichi.fujisawa.sb@hitachi.com.

0031-3203/$30.00 © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.03.015

engines have become ubiquitous and are expanding their reach into the areas of image documents, photographs, and videos. People are re-evaluating the importance of handwriting and trying to integrate it into the digital world. It seems that paper is still not going to disappear. Mobile devices with micro cameras now have CPUs capable of real-time recognition. The future prospects of these developments are discussed here.

2. Brief historical view

2.1. Overview

The first practical OCR appeared in the United States in the 1950s, in the same decade as the first commercial computer UNIVAC. Since then, each decade has seen advances in OCR technology. In the early 1960s, IBM produced their first models of optical readers, the IBM 1418 (1960) and IBM 1428 (1962), which were, respectively, capable of reading printed numerals and handprinted numerals. One of the models of those days could read 200 printed document fonts and were used as input apparatus for IBM 1401 computers. Also in the 1960s, postal operations were automated using mechanical letter sorters with OCRs, which for the first time automatically read postal codes to determine destinations. The United States Postal Service first introduced address-reading OCRs, which in 1965 began reading the city/state/ZIP line of printed envelopes [2]. In Japan, Toshiba and NEC developed handprinted numeral OCRs for postal code recognition, and put them into use in 1968 [3]. In Germany, a postal code system was introduced for the first time in the world in 1961 [4]. However, the first postal code reading letter sorter in Europe was introduced in Italy in 1973, and the first letter sorter with an automatic address reader was introduced in Germany in 1978 [5].

Japan started to introduce commercial OCRs in the late 1960s. Hitachi produced their first OCR for printed alphanumerics in 1968 and the first handprinted numeral OCR for business use in 1972. NEC developed the first OCR that could read handprinted Katakana in addition in 1976. The Japanese Ministry of International Trade and Industry (since renamed the Ministry of Economy, Trade and Industry) conducted a 10-year 20 billion-yen national project on pattern information processing starting in 1971. Among other research topics, Toshiba worked on printed Kanji recognition, and Fujitsu worked on handwritten character recognition. The ETL character databases including Kanji characters were created as part of this project, which contributed to research and development of Kanji OCRs [6]. As a by-product, the project attracted many students and researchers into the pattern recognition area. In the United States, IBM introduced a deposit processing system (IBM 3895) in 1977, which was able to recognize unconstrained handwritten check amounts. The author had a chance to observe it in operation at Mellon Bank in Pittsburgh in 1981, and it could reportedly read about 50% of handwritten checks with the remaining half being hand coded. The state of the art in character recognition in the 1960s and 1970s is well documented in the literature [7,8].

The 1980s witnessed significant technological advances in semiconductor devices such as CCD image sensors, microprocessors, dynamic random access memories (DRAMs), and custom-designed LSIs. For example, OCRs became smaller than ever fitting on desktops (Fig. 1). Then cheaper megabyte-size memories and CCD image sensors enabled whole-page images to be scanned into memory for further processing, in turn enabling more advanced recognition and wider applications. For example, handwritten numeral OCRs that could recognize touching characters were introduced for the first time in 1983, making it possible to relax physical form constraints and writing constraints. In the late 1980s, Japanese vendors of OCRs introduced into their product lines new OCRs that could recognize about 2400 printed and handprinted Kanji characters. These were used to read names and addresses for data entry. More detailed technology reviews are available in the literature [9,10].

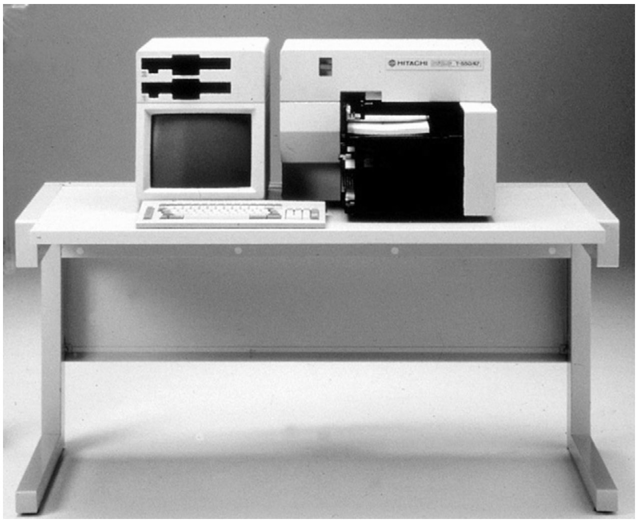

Fig. 1. First desktop Hitachi OCR HT-560 (1982).

The office automation boom of the 1980s, which was influential in Japan, had two features. One was Japanese language processing by computers and Japanese word processors. Emergence of Kanji OCRs was a natural consequence of this development. The other feature was optical disks used as computer storage systems, which were developed and put into use in the early 1980s. A typical application was patent automation systems in the United States and Japan that stored images of patent specification documents. The Japanese patent office system then stored approximately 50 million documents or 200 million digitized pages on 12-in optical disks. Each disk could store 7 GB of data, the equivalent of 200 000 digitized pages. The system used 80 Hitachi optical disk units and 80 optical library units. These systems can be considered one of the first digital libraries. This kind of new computer applications directly and indirectly encouraged studies on document understanding and document layout analysis in Japan. More importantly, it was in this decade that documents became the focus of computer processing for the first time.

The changes in the 1990s were due to the upgraded performances of UNIX workstations and then personal computers. Though scanning and image preprocessing were still done by the hardware, a major part of recognition was implemented by the software on general-purpose computers. The implication of this was that programming languages like c and c++ could be used to code recognition algorithms, allowing more engineers to develop more complicated algorithms and expanding the research community to include academia. During this decade, commercial software OCR packages running on PCs also appeared on the market. Techniques for recognizing freely handwritten characters were extensively studied, and successfully applied to bank check readers and postal address readers. Advanced layout analysis techniques enabled recognition of wider varieties of business forms. Research institutions specializing in this field such as CENPARMI, led by Prof. Suen and CEDAR, led by Prof. Srihari and Prof. Govindaraju contributed to these advances. New high-tech vendors appeared, including A2iA, which was started by the late Prof. Simon in France [11], and Parascript, which was started in Russia to do business in the United States. In Japan, the Japanese Postal Ministry conducted the third generation postal automation project between 1994 and 1996, in which Toshiba, NEC, and Hitachi joined to develop postal address recognition systems that could sort sequences. This project enabled significant advances in Japanese address reading.