The Evolution of Document Image Analysis

3

Henry S. Baird and Karl Tombre

Contents

Abstract

One of the first application domains for computer science was Optical Character Recognition. At that time, it was expected that a machine would quickly be able to read any document. History has proven that the task was more difficult than that. This chapter explores the history of the document analysis and recognition domain, from OCR to page analysis and on to the open problems which are still to be completely dealt with.

Keywords

Document image analysis • History • OCR

H.S. Baird
Computer Science and Engineering Department, Lehigh University, Bethlehem, PA, USA
e-mail: baird@cse.lehigh.edu

K. Tombre
Université de Lorraine, Nancy, France
e-mail: Karl.Tombre@loria.fr; Karl.Tombre@univ-lorraine.fr

D. Doermann, K. Tombre (eds.), *Handbook of Document Image Processing and Recognition*, DOI 10.1007/978-0-85729-859-1_43,

Introduction

The first computers and computer science as a field emerged from World War II. Once this field extended beyond scientific computing and defense applications, one of its first uses was Optical Character Recognition (OCR). At that time, it was expected that a machine would soon be able to read any document. But it quickly became evident that progress would be slower than expected and that the vast diversity of applications and needs would make it impossible to just rely on improved scores for single-character recognition. Every error in a single digit for a ZIP code can send a letter to the wrong destination, five character errors on a page of text are often enough to add a significant cost of post-OCR editing, and when the problem comes to extracting the major information relayed by a complete document, without knowing the typewriting font or when it is handwritten, the challenge becomes much more one of analyzing the global document "scene" than to decipher single characters.

This vast diversity of document analysis problems, almost none of which could be automatically analyzed with the state of the art, led to extreme specialization of the systems and to a slow pace of development for new applications [1].

This chapter does not pretend to provide a complete history of how document analysis systems evolved, but points out some of the main landmarks in this evolution.

Isolated Character Recognition

Schantz, in his history of OCR [1], mentions an 1809 US patent for assisting reading by the blind. By 1870, a mosaic of photocells was used by C. R. Carey to transmit images, and by the turn of the 20th century, P. Nipkow used "scanning disk" to analyze images line by line. In 1912 Emmanuel Goldberg patented a machine to convert typed messages to telegraphic messages, and in 1914 Edmund F. D'Albe invented a handheld scanner that converts images of text into sounds to assist blind readers. By 1931 Goldberg patented an optical scanner driving a template-matching character classifier. Accuracy of this and many similar technologies was limited, until the 1950s, by "difficulties in precisely aligning source patterns with template patterns."

In 1951 David Shepard demonstrated an OCR machine able to read 23 letters typed by a "standard typewriter." Around the same time Jacob Rabinow refined template matching to search among a range misregistrations etc. for a "best match" and extended the alphabet to both upper and lower case. In the early 1950s advances seemed to depend, principally, on better imaging sensors and controlling movement of sensors and paper. Recognition methods included "area correlation, feature analysis, [and] curve tracing"; refinements included "noise filtering [and] image preprocessing." Applications spread rapidly in the late fifties; most were custom

designed: in 1959, a machine built for the US Air Force could read both upper and lower case alphanumerics – but only in a single typeface. Even typewriter fonts posed a challenge in 1954 an OCR manufacturer proposed widespread standardization of printing using a specially designed 5×7 grid typeface, in order to assist OCR.

In the 1960s typewriting applications spread rapidly even though each was typically custom trained for the application and could handle only one typeface; still, it was generally believed that the technology could be trained to usefully high accuracy on "almost any consistently formed character patterns." OCR machines could read only a single format of document in a batch: users were expected to standardize their input. Soon customers were asking for machines that could handle a variety of documents including a variety of typefaces. The first commercial "multifont" machine appeared in 1964; by the late 1960s such a machine, custom built for the US Army, "read 63 % [. . ..] error-free" of a highly heterogeneous input stream of documents containing upper and lower case letters in over 30 pretrained fonts. Through the 1960s all OCR machines were large custom installations handling large batches, often at very high speeds.

The early 1960s saw the first promising experiments on hand-printed (not cursive) characters. In the late 1960s, user-trainable OCR machines appeared, which were then marketed as (potentially) "omnifont"; but these were seldom effective. The technical obstacles facing the technology are reflected clearly in this proposal by Rabinow in [2]:

> The more control one puts into a document, the simpler and less costly is the reading machine. . . . [How can this be done?] Standardize the type of paper, [. . .] the size of paper, [. . .] the quality of printing, [. . .] the format, and [. . .] the font.

The industry echoed this cry, and the two standard OCR fonts (OCR-A and OCR-B) resulted.

By the late 1970s, character readers were complemented by other control-intensive technologies including bar code readers and mark sense readers. The early 1980s saw a significant change in the market as fax machines spread, and higher-resolution flat-bed document scanners became affordable, and these were connected to personal computers. Then, OCR companies raced to deliver "personal OCR" where pre-training and most kinds of "control" were missing. Through massive training on large collections (of tens of millions) of character images from scores of typefaces, OCR companies tried to fulfill their claim of truly "omnifont" recognition systems.

In 1992, Mori et al. presented a historical review of these early years in OCR Research and Development [3]. At the same time, George Nagy [4] prophetically criticized as "exhausted" the then state of the art which relied on accurate recognition of isolated character images, and pointed to the promise of exploiting larger contexts including style consistency within documents, and broader multicharacter contextual analysis including layout context. Mori, Nishida, and Yamada [5] later summarized the state of the art in 1999 of approaches to isolated character recognition.

Beyond Recognition of Isolated Characters: Exploitation of Context

As pointed out by Nagy, it was necessary to go beyond progress in recognition rates for isolated characters if machines were expected to read like human beings do. As a matter of fact, we humans do not only learn the individual letters, in first grade; we also learn to read and understand complete texts, to extract meaningful information from forms, and to communicate with each other through sophisticated documents such as accountancy reports, news articles, poetry, or even maps and engineering drawings. Even handwritten documents where any single character may be extremely hard to decipher become meaningful because we take the context of the document into account.

This starts by looking at a typewritten, printed, or handwritten document like a message using a *language* known to both emitter and receiver. Thus, the analysis of a written objet can take into account linguistic aspects.

Sampson, in his seminal study Writing Systems [6] noted that

> although the tide is beginning to turn now [1985], for most of the twentieth century linguistics has almost wholly ignored writing.

Thus, serious academic attention to the linguistics of writing is very recent, contemporary with the appearance of OCR machines designed for "general-purpose" use by nonexperts. Within the academic linguistics community, computational approaches have also been very much in the minority through the 1970s; even today, it can be difficult for an OCR researcher to locate counterparts in the linguistics community who are willing and able to share their insights in readily usable forms of data and software. We believe that this fact, in turn, has significantly slowed the exploitation of knowledge discovered by linguists with document image recognition technology.

One of the earliest steps in automating natural language text is to provide a means for checking the legality of words. The simplest such means is surely a more or less exhaustive list, or flat lexicon. The earliest exploitation of dictionary context in OCR systems relied on such lists and continues to do so up to the present. The collection of computer-legible lexica accelerated rapidly in the 1970s and is now nearing saturation for languages supported by a robust information technology industry; but, as always, many remote languages are underserved, and thus the extension of modern OCR systems, designed to be cheaply retargeted to new languages by providing a lexicon, may hit a significant barrier.

Many languages however are highly *inflected* so that a large number of variations of words occur: the characteristics common to them all is sometimes called the stem word, and the variations are often provided by suffixes, prefixes, and more complex rewritings. Latin, Spanish, and Russian are extreme cases. For most of these languages, it is possible to capture all or most of the *inflectional morphology* rule within a computational linguistics algorithm, which offers several benefits:

1. **Smaller lexicons**, since many variations collapse into the same rule.
2. **Ease of entering new words**, since with the addition of only the stem of a new word, all its inflections are covered.

3. **Recognition of neologisms**, so that words never seen before can be correctly identified (by derivational morphology).
4. **Faster lookup** is a possibility, in spite of computational overhead, in cases where an equivalent lexicon is unmanageably huge.

All of these benefits are potentially enjoyed by an OCR system.

Ritchie et al. [7] presented a nearly exhaustive analysis of such morphological structure of English words, which requires a "two-level" system of regular grammar rewriting rules. They mention that this approach works on languages including Finnish, French, German, Japanese, Rumanian, Old Church Slavonic, and Swedish. Unfortunately, Semitic languages such as Hebrew and Arabic possess "non-concatenative" morphologies which require more advanced models. An implication for OCR systems is that linguistic contexts as elementary as "dictionary checks" may not be feasible even today for underserved languages, and making progress may require professional linguistic effort or even research in linguistics.

From Words to Pages, from Pages to Structured Documents, and Onwards to Non-textual Documents

One thing is to work on recognizing characters, words, or sentences; another is to get access to all the information present in a document such as a letter to be handled by the postal service, a bank check, a completed form, or a business letter. Beyond character and word recognition, this includes numerous tasks, especially related to the spatial analysis of the document page, which is actually a scene analysis problem, and the mapping between the layout structure and the semantics conveyed by this layout.

Early work in that area addressed the most common layouts. The layout in rectangular shapes, which can be found in books, newspapers, journals, etc., was extracted through various methods designed by research teams in the 1980s. Methods such as the run length smoothing algorithm designed at IBM [8] and used in systems analyzing newspaper archives [9], or the X-Y tree which decomposed a journal article into homogeneous parts [10], are still widely used nowadays, as explained in ▸Chap. 5 (Page Segmentation Techniques in Document Analysis).

Specific classes of documents, where the layout and/or syntactical constraints are strong and well known, and the need for reliability on large volumes of documents is high, were also given special attention very early. Thus, systems were designed for postal automation [11] or bank check recognition (see ▸Chap. 21 (Document Analysis in Postal Applications and Check Processing) for a history of this field), tables and forms (see ▸Chap. 19 (Recognition of Tables and Forms)), or business letters.

It also became necessary to look beyond text, as documents in the most general meaning are formalized ways for humans to communicate with each other, using a commonly understood language, which can also include graphical parts, images, etc. This led to work on systems for the analysis of maps [12, 13], of electrical diagrams [14], or of engineering drawings [15]. If these early systems were often limited, fine-tuned for a narrow set of documents, and difficult to maintain and to expand,

they still helped in developing basic methods for graphics recognition which are still in use, as detailed in ▸Chaps. 15 (Graphics Recognition Techniques), ▸16 (An Overview of Symbol Recognition), and ▸17 (Analysis and Interpretation of Graphical Documents).

Stubborn Obstacles to Document Image Recognition

In 1982, Schantz said that "the rate of correct character recognition is directly proportional to the quality of source data" [1]. In 1999, Rice, Nagy, and Nartker [16] published a profusely illustrated taxonomy of frequently occurring OCR errors and discussed the origins of these errors with unprecedented insight. One aspect of quality, which turns out to be amenable to formal modeling and systematic engineering attack, is the degradation of the image due to both printing and image capture [17].

Among the many obstacles which are still on the road of the evolution of document image analysis and recognition, let us mention those which seem to us to be the most difficult to deal with and which thus have to be given continuing attention in the coming years.[1]

(a) Document images are not always captured in an optimal and controlled way, and their quality is often too low. In some cases, such as managing large collections of heritage documents, decisions can have been made on the resolution of the scanning process, and the documents themselves are sometimes degraded. Later processes have to use the images as they are, even when it becomes evident that the quality is far from being adapted to the analysis processes. Other cases where the image quality can lead to specific problems include text in video and documents captured by a camera or a phone (see ▸Chap. 25 (Text Localization and Recognition in Images and Video)).

(b) Many recognition processes rely on classification methods which need to be trained. But it is not always possible to work with large enough sets of training samples, covering the full diversity of the analysis problem. This is especially true for non-textual documents. Related to that, for the purpose of evaluating the performances of document analysis systems, it is often difficult to have sufficiently ground-truthed data. See ▸Chaps. 29 (Datasets and Annotations for Document Analysis and Recognition) and ▸30 (Tools and Metrics for Document Analysis Systems Evaluation) for further discussions of these problems.

(c) We have seen that linguistic tools are an important asset in designing efficient document analysis systems. But in many languages, such tools lack or are not sufficiently developed.

(d) No document analysis system can be made completely automatic, so that it could just work as a post-processing step on the output of the scanner. But it is difficult to construct effective user interfaces, to effectively integrate document image analysis into a larger workflow, even the more because it is not

always easy to have the user accept the error-prone nature of document image processing and recognition.

(e) There seems to be infinite ways in which people create documents, with complex layouts or with inconsistent or nonexisting typographical and semantic rules. It is impossible to train a system for all such variations. This has led researchers and companies to focus on small subsets of problems, and the solutions they design are very often not applicable to slightly different problems or document categories.

(f) Today, many companies are faced with the problem that their customers or suppliers send documents via multiple channels, in order to communicate messages having a legal or economic relevance. This includes filled forms via printed mail, faxes, scanned document images sent as email, or even electronic documents in PDF or TIFF format, complemented by metadata. Although most of these channels provide a certain amount of metadata (a fax provides a fax number, emails have information in their headers, and electronic documents have full sets of descriptors), they still necessitate addressing the whole range of document analysis problems, as this book largely demonstrates. Moreover, the messages conveyed by the document are an integral part of the workflow, i.e., they may request information or answer such a request. Helping to make this multichannel information directly available to feed the workflow is a challenge, and good solutions to this challenge would have a high economic value.

Conclusion

Sellen and Harper [18] cogently argue that paper's role as a medium for communication is not likely to decline in scale within the foreseeable future, even as purely digital media continue to grow exponentially. Lesk's prophetic study of digital libraries [19] points out that, even as inevitably much modern data will be "born digital" and so never have to be converted from images of documents, the total volume of printed paper will grow alongside the accelerating scale of digital libraries. In Nunberg's vision of the future of the book [20], digital and document-based versions of information will coexist and, assisted by document image analysis technology, refer richly to one another.

Cross-References

▸An Overview of Symbol Recognition
▸Analysis and Interpretation of Graphical Documents
▸Analysis and Recognition of Music Scores
▸Analysis of the Logical Layout of Documents
▸Document Analysis in Postal Applications and Check Processing
▸Graphics Recognition Techniques
▸Logo and Trademark Recognition

▸Machine-Printed Character Recognition
▸Page Segmentation Techniques in Document Analysis
▸Processing Mathematical Notation
▸Recognition of Tables and Forms

Notes

[1]Many thanks to the authors of several chapters of this handbook for their contribution to the list of stubborn obstacles.

References

1. Schantz HF (1982) History of OCR, optical character recognition. Recognition Technologies Users Association, Manchester Center, Vt., USA
2. Rabinow J (1969) Whither OCR? And whence? Datamation 15(7):38–42
3. Mori S, Suen CY, Yamamoto K (1992) Historical review of OCR research and development. Proc IEEE 80(7):1029–1058
4. Nagy G (1992) At the frontiers of OCR. Proc IEEE 80(7):1093–1100
5. Mori S, Nishida H, Yamada H (1999) Optical character recognition. Wiley, New York
6. Sampson G (1985) Writing systems. Stanford University Press, Stanford
7. Ritchie G, Russell G, Black A, Pulman S (1992) Computational morphology. MIT, Cambridge
8. Wong KY, Casey RG, Wahl FM (1982) Document analysis system. IBM J Res Dev 26(6): 647–656
9. Wang D, Srihari SN (1989) Classification of newspaper image blocks using texture analysis. Comput Vis Graph Image Process 47:327–352
10. Nagy G, Seth S, Viswanathan M (1992) A prototype document image analysis system for technical journals. IEEE Comput Mag 25(7):10–22
11. Schürmann J (1978) A multifont word recognition system for postal address reading. IEEE Trans Comput 27(8):721–732
12. Kasturi R, Alemany J (1988) Information extraction from images of paper-based maps. IEEE Trans Softw Eng 14(5):671–675
13. Shimotsuji S, Hori O, Asano M, Suzuki K, Hoshino F, Ishii T (1992) A robust recognition system for a drawing superimposed on a map. IEEE Comput Mag 25(7):56–59
14. Groen F, Sanderson A, Schlag J (1985) Symbol recognition in electrical diagrams using probabilistic graph matching. Pattern Recognit Lett 3:343–350
15. Vaxivière P, Tombre K (1992) Celesstin: CAD conversion of mechanical drawings. IEEE Comput Mag 25(7):46–54
16. Rice S, Nagy G, Nartker T (1999) OCR: an illustrated guide to the frontier. Kluwer, Boston
17. Baird H (2007) The state of the art of document image degradation modeling. In: Chaudhuri B (ed) Digital document processing. Springer, London
18. Sellen A, Harper R (2003) The myth of the paperless office. MIT, Cambridge
19. Lesk M (1997) Practical digital libraries: books, bytes, & bucks. Morgan Kaufmann, San Francisco
20. Nunberg G (1996) The future of the book. University of California Press, Berkeley
21. Baird HS, Bunke H, Yamamoto K (eds) (1992) Structured document image analysis. Springer, Berlin/New York
22. Nagy G (2000) Twenty years of document image analysis in PAMI. IEEE Trans PAMI 22(1):38–62

Further Reading

The beginning of the 1990s was a time when document analysis had already matured substantially, and the research community felt the need for a consolidation of the knowledge gained that far. Several special issues and books were published in that period and are a good start to understand the early years of document image processing and recognition.

- The Proceedings of the IEEE, vol. 80, no. 7, 1992, contained several interesting articles on the history and the state of the art, including Refs. [5] and [4].
- A special issue of IEEE COMPUTER Magazine (vol. 25, no. 7, July 1992) had a number of articles on state of the art document analysis systems, including (but not limited to) [13] and [15].

The early 1990s was also a time when the document image analysis community created its mainstream conference, the International Conference on Document Analysis and Recognition, whose first instance was held in Saint-Malo, France, in 1991. Prior to that, a workshop in New Jersey held in 1990 gave way to a book which also gave a good overview of the state of the art of the field at that time [21].

In 2000, George Nagy published a survey article revisiting 20 years of document image analysis seen through the window of the IEEE Transactions on Pattern Analysis and Machine Intelligence [22].

For review of later evolutions, the reader is referred to the "Further readings" of the different chapters of this book, as the field became very broad, with specializations on many different topics.