OCR: Print - An overview

Abdel Belaïd

CRIN-CNRS
Campus Scientifique, B.P. 239
54506 Vandœuvre-lès-Nancy Cedex
France
email: abelaid@loria.fr

1 Introduction

Nowadays, there is much motivation to provide computerized document analysis systems. Giant steps have been made in the last decade, both in terms of technological supports and in software products. Character recognition (OCR) contributes to this progress by providing techniques to convert large volumes of data automatically. There are so many papers and patents advertising recognition rates as high as 99.99 percent; this gives the impression that automation problems seem to have been solved. However, the failure of some real applications show that performance problems subsist on composite and degraded documents (i.e. noisy characters, tilt, mixing of fonts, etc.) and that there is still room for progress. Various methods have been proposed to increase the accuracy of optical character recognizers. In fact, at various research laboratories, the challenge is to develop robust methods that remove as much as possible the typographical and noise restrictions while maintaining rates similar to those provided by limited-font commercial machines.

There is a parallel analogy between the various stages of evolution of OCR systems and those of pattern recognition. To overcome the recognition deficiency, the classical approach focussing on isolated characters has been replaced with more contextual techniques. The opening of OCR domain to document recognition leads to combination of many strategies such as document layout handling, dictionary checking, font identification, word recognition, integration of several recognition approaches with consensual voting, etc.

The rest of this paper is devoted to a summary of the state of the art in the domain of printed OCR (similar to the presentations in [Imp 91, Gov 90, Nad 84, Man 86]), by focussing attention essentially on the new orientations of OCR in the document recognition area.

2 Document Analysis Aspects

Characters are arranged in document lines following some typesetting conventions which we can use to locate characters and find their style. Typesetting rules can help in distinguishing such characters as "s" from "5", "h" from "n" and "g" from "9" which can be often confused in multifont context [Kah 87]. They can also limit the search area according to characters relative positions and heights with respect to the baseline [Luc 91a, Luc 91b, Kan 90]. The role of typesetting cues to aid document understanding is discussed by Holstege and Tokuda [Hol 91].

2.1 Layout Segmentation

Location of characters in a document is always preceded by a layout analysis of the document image. The layout analysis involves several operations such as determining the skew, separating picture from text, and partitioning the text into columns, lines, words, and connected components. The portioning of text is effected through a process known as segmentation. A survey of segmentation techniques is given in [Nad 84].

2.2 Character Building

In building character images, one is often confronted with touching or broken characters that occur in degraded documents (such as fax, photocopy, etc.). It is still challenging to develop techniques for properly segmentating words into their characters. Kahan et al. [Kah 87] detected touching characters by evaluation of vertical pixel projection. They executed a branch-and-bound search of alternative splittings and merges of symbols pruned by word-confidence scores derived from symbol confidence. Tsujimoto and Asada [Tsu 91] used a decision tree for resolving ambiguities. Casey and Nagy [Cas 82] proposed a recursive segmentation algorithm. Liang et al. [Lia 93] added to this algorithm, contextual information and a spelling checker to correct errors caused by incorrect segmentation. Bayer [Bay 87] proposed a hypothesis approach for merging and splitting characters. The hypotheses are tested by several experts to see whether they represent a valid character. The search is controlled by the A* algorithm resolving backtracking processing. The experts comprise the character classifier and a set of algorithms for context processing.

2.3 Font Consideration

A document reader must cope with many sources of variations notably that of font and size of the text. In commercial devices, the multifont aspect was for a long time neglected for the benefit of speed and accuracy, and substitution solutions were proposed. At first, to cater for some institutions, the solution was to work on customized fonts (such as OCR-A and OCR-B) or on a selected font from a trained library to minimize the confusion between similar looking characters. The accuracy was quite good, even on degraded images on the condition that the font is carefully selected. However,

recognition scores drop rapidly when fonts or sizes are changed. This is due to the fact that the limitation to one font naturally promotes the use some simple and sensitive pattern recognition algorithms, such as template matching [Dud 73].

In parallel with commercial investigations, the literature proposed multifont recognition systems that are based on typographical features. Font information is inherent in the constituent characters [Rub 88] and feature-based methods are less font sensitive [Sri 84, **?**, Kah 87]. Two research paths were taken with multifont machines. One gears towards the office environment. This introduced systems which can be trained by the user to read any given font [Sch 78, Shl 88, Bel 91, Ani 91a, Ani 91b]. The system is only able to recognize a font from among those learned. The others try to be font independent. The training is based on pattern differentiation rather than on font differentiation [Lam 87, Bai 86, Bai 91].

3 Character Recognition

3.1 Feature Extraction

This step is crucial in the context of document analysis where several variations may be caused by a number of different sources: geometrical transformation because of low data quality, slant and stroke width variation because of font changing, etc. It seems reasonable to look for features which are invariant and which capture the characteristics of the character by filtering out all attributes which make the same character assume different appearances. The classifier could store a single prototype per character. Schurmann et al.[Sch 92] applies normalizing transformations to reduce certain well-defined variations as far as possible. The inevitably remaining variations are left for learning by statistical adaptation of the classifier.

3.2 Character Learning

The keys of printed character learning are essentially training set and classification adaptation to new characters and new fonts. The training set can be given either by user or extracted directly from document samples. In the first case, the user selects the fonts and the samples to represent each character in each font and then guides the system to create models as in Anigbogu [Ani 91b]. Here, the user must use sufficient number of samples in each font according to the difficulty of its recognition. However, it is difficult in an omnifont context to collect a training set of characters having the expected distribution of noise and pitch size. Baird [Bai 90] suggested parameterized models for imaging defects, based on a variety of theoretical arguments and empirical evidence. In the second case, the idea is to generate the training set directly from document images chosen from a wide variety of fonts and image quality and to reflect the variability expected by the system [Bok 92]. The problem here is that one is not sure that all valid characters are present.

3.3 Contextual Processing

Contextual processing attempts to overcome the short coming of decisions made on the basis of local properties and to extend the perception on relationships between characters into word. Most of the techniques try to combine geometric information, as well as linguistic information. See [Sri 85] for an overview of these techniques. Anigbogu and Belaïd [Ani 91a, Ani 91b, Bel 91] used Hidden Markov Models for character and word modeling. Characters are merged into groups which are matched against words in a dictionary using Ratcliff/Obershelp pattern matching method. In the situation where no acceptable words are found, the list of confused characters is passed through a Viterbi net and the output is taken as the most likely word. The bigram and character position-dependent probabilities used for this purpose were constructed from a French dictionary of some 190,000 words. The word-level recognition stands at over 98%.

4 Commercial Products

Commercial OCR machines came in practically at the beginning of 1950's and have evolved in parallel with research investigations. The first series of products heavily relied on customized fonts, good printing quality and very restricted document layout. Nowadays, we can find a vast range of products, more powerful than the previous ones. Among these are, certain hand-held scanners, page readers, integrated flat-bed and document readers. The tendency is to use the fax machine as an image sensor. Instead of printing the fax message on paper, it is taken directly as input to OCR system. It is to be noted that the obtained images are of a poor quality. The challenge in this area is the development of high performing tools to treat degraded text with results as good as those of classical OCR's.

OCR is used in three main domains: the banking environment, for data entry and checking, office automation, for text entry and the post office for mail sorting. We can find many surveys on commercial products in [Mor 92, Man 86, Bok 92, Nag 92]. Recently, the Information Science Research Institute had the charge to test technologies for OCR from machine printed documents. A complete review appeared giving a benchmark of different products in use in the U.S. Market.

5 Conclusion

We have attempted to show that OCR is an essential part of the document analysis domain. Character recognition cannot be achieved without typesetting cues to help the segmentation in a multifont environment. We have also shown the unavoidable recourse to linguistic context; the analysis must be extended to this domain. The training still remains the weak side of OCR for now for it is difficult to generate a training set of characters which includes all the variability the system will be expected to handle. Finally, it would appear more and more that in real-world OCR many dif-

ferent techniques must be combined to yield high recognition scores [Ani 91b, Ho 92]. For this reason, the tendency is to combine the results of many OCR systems in order to obtain the best performance possible.

References

[Ani 91a] J. C. Anigbogu and A. Belaïd. Application of Hidden Markov Models to Multifont Text Recognition. In *First International Conference on Document Analysis and Recognition (ICDAR'91)*, volume II, pages 785–793, St-Malo, France, September 1991.

[Ani 91b] J. C. Anigbogu and A. Belaïd. Recognition of Multifont Text Using Markov Models. In 7^{th} *Scandinavian Conference on Image Analysis*, volume I, pages 469–476, August 1991.

[Bai 86] H. S. Baird, S. Kahan, and T. Pavlidis. Components of a Omnifont Page Reader. In *Proceedings of 8th International Conference on Pattern Recognition*, pages 344–348, Paris, 1986.

[Bai 90] H. S. Baird. Document Image Defect Models. In *Proceedings of the Workshop on Syntactical and Structural Pattern Recognition*, pages 38–47, 1990.

[Bai 91] H. S. Baird and R. Fossey. A 100-font Classifier. In *First International Conference on Document Analysis and Recognition (ICDAR'91)*, volume 1, pages 332–340. IAPR, 30 Sept. - 2 Oct. 1991.

[Bay 87] T. Bayer. Segmentation of Merged Character Patterns with Artificial Intelligence Techniques. In *Proceedings of 5th Scandinavian Conference on Image Analysis*, pages 49–55, Stockholm, 1987.

[Bel 91] A. Belaïd and J. C. Anigbogu. Text Recognition Using Stochastic Models. In R. Gutiérrez and M. J. Valderrama, editors, 5^{th} *International Symposium on ASMDA*, pages 87–98. World Scientific, April 1991.

[Bok 92] M. Bokser. Omnidocument technologies. *Proceedings of the IEEE*, 80(7), July 1992.

[Cas 82] R. G. Casey, T. D. Friedman, and K. Y. Wong. Automatic Scaling of Digital Print Fonts. *IBM Journal of Research and Development*, 26(6): 657–666, 1982.

[Dud 73] R. Duda and P. E. Hart. *Pattern Recognition and Scene Analysis.* John Wiley and sons, New York, 1973.

[Gov 90] V. K. Govindan and A. P. Shivaprasad. Character Recognition – a Review. *Pattern Recognition*, 23(7): 671–683, 1990.

[Ho 92] T. K. Ho. *A Theory of Multiple Classifier Systems and Its Application to Visual Word Recognition.* PhD thesis, State University of New York at Buffalo, May 1992.

[Hol 91] M. Holstege, Y. J. Inn, and L. Tokuda. Visual Parsing: An Aid to Text Understanding. In *Recherche d'Information Assistée par Ordinateur*, volume RIAO-91, pages 175–193, 1991.

[Hul 83] J. J. Hull, S. N. Srihari, and R. Choudhari. An Integrated Algorithm for Text Recognition: Comparison with a Cascaded Algorithm. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, PAMI-5(4): 384–395, July 1983.

[Imp 91] S. Impedovo, L. Ottaviano, and S. Occhinegro. Optical Character Recognition - A Survey. *International Journal of Pattern Recognition and Artificial Intelligence*, 1&2(5): 1–24, 1991.

[Kah 87] S. Kahan, T. Pavlidis, and H. S. Baird. On the Recognition of Printed Characters of Any Font and Size. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 9(2): 274–288, 1987.

[Kan 90] J. Kanai. Text Line Extraction and Baseline Detection. In *Recherche d'Information Assistée par Ordinateur*, volume RIAO-90, pages 194–209, 1990.

[Lam 87] S. W. Lam and H. S. Baird. Performance Testing of Mixed-font Variable-size Character Recognizers. In *Proceedings of 5th Scandinavian Conference on Image Analysis*, pages 563–570, Stockholm, 1987.

[Lia 93] S. Liang, M. Ahmadi, and M. Shridar. Segmentation of Touching Characters in Printed Document Recognition. In *Second International Conference on Document Analysis and Recognition (ICDAR'93)*, pages 569–572, Tsukuba, Japan, 1993.

[Luc 91a] P. G. De Luca and A. Gisotti. How to Take Advantage of Word Structure in Printed Character Recognition. In *Recherche d'Information Assistée par Ordinateur*, volume RIAO-91, pages 148–159, 1991.

[Luc 91b] P. G. De Luca and A. Gisotti. Printed Character Preclassification Based on Word Structure. *Pattern Recognition*, 24(7): 609–615, 1991.

[Man 86] J. Mantas. An Overview of Character Recognition Methodologies. *Pattern Recognition*, 19(6): 425–430, 1986.

[Mor 92] S. Mori, C. Y. Suen, and K. Yamamoto. Historical review of ocr research and development. *Proceedings of the IEEE*, 80(7): 1029–1058, July 1992.

[Nad 84] M. Nadler. A Survey of Document Segmentation and Coding Techniques. *Computer Vision, Graphics, and Image Processing*, 28: 240–262, 1984.

[Nag 92] G. Nagy. At the frontiers of OCR. *Proceedings of the IEEE*, 80(7): 1093–1100, July 1992.

[Rub 88] R. Rubinstein. *Digital Typography : An Introduction to Type and Composition for Computer System Design.* Addison Wesley, 1988.

[Sch 78] J. Schürmann. A Multifont Word Recognition System for Postal Address Reading. *IEEE Transactions on Computers*, C-27(8): 721–732, 1978.

[Sch 92] J. Schürmann, N. Bartneck, T. Bayer, J. Franke, E. Mandler, and M. Oberlander. Document Analysis - From Pixels to Contents. *Proceedings of the IEEE*, 80(7): 1101–1119, July 1992.

[Shl 88] S. Shlien. Multifont Character Recognition For Typeset Documents. *International Journal of Pattern Recognition and Artificial Intelligence*, 2(4): 603–620, 1988.

[Sri 84] S. N. Srihari. *Computer Text Recognition and Error Correction.* IEEE Computer Society Press, Silver Spring, MD, 1984.

[Sri 85] S. N. Srihari, J. J. Hull, et al. Adress Recognition Techniques in Mail Sorting: Research Directions. Technical Report 85-09, Department of Computer Science, SUNY, Buffalo, 1985.

[Tsu 91] Y. Tsujimoto and H. Asada. Resolving Ambiguity in Segmenting Touching Characters. In *First International Conference on Document Analysis and Recognition (ICDAR'91)*, pages 701–709, St-Malo, France, November 1991.