MVA '96 IAPR Workshop on Machine Vision Applications, November. 12-14, 1996, Tokyo, Japan

IDA: A System for Automated Sorting, Indexing, and Classification of Documents

Gerd Maderlechner, Thomas Brückner, and Peter Suda *
Corporate Research and Technology
Siemens AG

Abstract

IDA (Intelligent Document Analysis) is a modular software system, which assists to automate paper document entry. IDA consists of the following components: layout analysis, preclassification, OCR interface, fuzzy string matching, text categorization, lexical, syntactical and semantic analysis. The system has been applied to a variety of tasks: Presorting of forms, reports and letters, index extraction for archiving and retrieval, text column analysis in real estate register documents, in-house mail distribution, and classification of business letters by text content. This paper presents an overview of the architecture and applications of the system.

1 Introduction

The use of electronic images instead of paper in the office work flow has proven to provide a high potential for office automation and considerable increase in productivity. The bottlenecks in present document image processing systems are the manual sorting and indexing steps necessary for efficient document management and information retrieval.
Considerable research and development has been invested into document analysis and OCR by academia and industry, in particular for specific applications like forms processing, business letter interpretation, mail sorting etc., see e.g. [2], [3], [4], [7]. The design goal of IDA (Intelligent Document Analysis) is to cope with a large variety of printed documents without loosing high performance. The basic approach is to use both geometric (layout) and textual (content) knowledge of the documents, and to apply OCR only if necessary. IDA has no OCR of its own but relies on the OCR systems available on the market. In many applications the accuracy of OCR is sufficient for machine printed text if there is an adequate rejection strategy. Speed of OCR is a limiting factor for IDA running on standard hardware. Therefore focus of attention techniques have been developed to limit OCR evaluation to few regions of interest.

*Address: Otto-Hahn-Ring 6, D-81730 München, Germany. E-mail: gm@zfe.siemens.de

2 Architecture of the system

IDA has a modular and open architecture with three main components, layout analysis, OCR, and content analysis. The layout analysis consists of the following modules (Fig. 1):

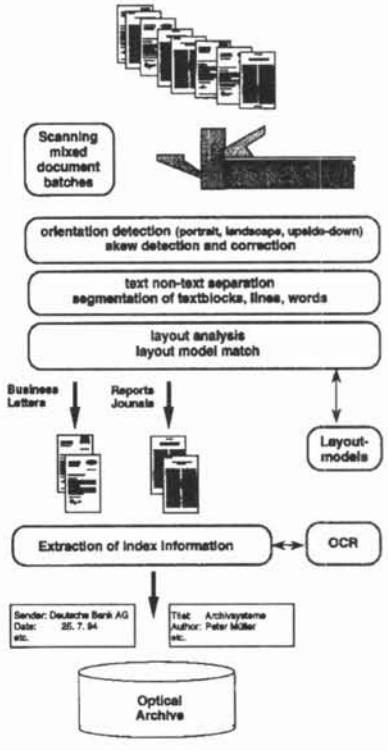

Figure 1: Overview of the IDA (Intelligent Document Analysis) system with modules for document sorting and indexing.

• Fast orientation (portrait vs. landscape) and upside-down detection. • Fast skew detection and correction • Configurable layout segmentation based on both projections and connected components al-

lowing a simultaneous top-down and bottom-up approach. • Model-based document analysis by using specific document layout models. • Interactive model acquisition tool based on layout segmentation results to generate reference models for different types of documents, e.g. business letters and reports. • A fast algorithm for layout pattern matching, i.e. comparison of a given document with the reference models.

• IDA provides a flexible OCR interface for different OCR engines using applications programmers interfaces (API).

The content analysis consists of the following modules (Fig. 2):

• Morphological analysis by lemmatizer for reduction to word stems. • Tagger for labeling the correct syntactic word category. • Training module for building models of text categories using labeled text classes. • Classification into different text categories by similarity match with models.

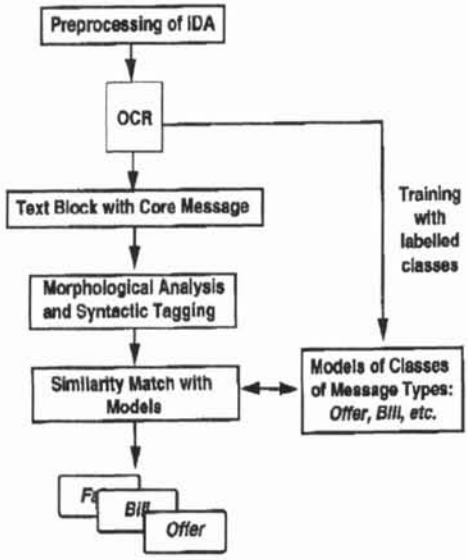

Figure 2: IDA modules for content analysis and text classification.

3 Layout Analysis

The high performance of layout analysis is achieved by efficient representations of the binary image using runlength code (RLC), connected components (CC) and tiling for fast random access.

Preprocessing The features for orientation detection are calculated on the fly during reading the image file. After eventual n*90 degree rotation into correct reading orientation the skew detection is done in two steps from coarse (about 5 degrees) to fine values (about 0.2 degrees), using text line features. This correction is performed by a fast rotation module.

Document Image Segmentation Document segmentation divides the page into rectangular text and non-text blocks. IDA uses both a top down approach based on white stream analysis [6] and a bottom up aggregation of CCs and text line portions to build rectangular text blocks. Text blocks contain text lines and are attributed with information on line spacing, font size, text alignment, etc. The remaining rectangular blocks are labeled as non-text blocks, which are candidates for logos and figures. Horizontal and vertical lines are determined using local projections and histograms, described in [8].

Layout Model Matching Most legal and business documents have a characteristic layout. Forms are dominated by tabular structures, boxes, and preprinted text. Business letters contain at least a logo at top, and an information on the sending organization. This preprinted information is used as reference layout model information. To provide the system with models there is an interactive acquisition module that takes a segmented image and asks the user to select the preprinted logo, sender address, etc. The geometric information and the corresponding textual information are stored in a model base.

In the recognition phase the system matches the layout regions of the input document against all reference models. The matching quality is calculated by the size and relative positions of all corresponding regions, allowing translation and small distortions.

4 OCR

In the PC (windows) world there is a larger choice of OCR systems available than in the workstation (Unix) domain. We transferred IDA from Sun to PC platform and got higher performance for IDA as well as for OCR on a 100 MHz PC system. We tested five OCR engines (from Caere, Calera, Cognitive, Expervision, and Xerox Imaging Systems). The applications programmer interfaces of these OCR systems allow the access to regions of interest in the document image. The IDA interface offers a flexible representation of the OCR results, including alternatives and confidence values, depending on the OCR. The respective accuracies of the OCR systems correspond fairly well with the results published by ISRI [5]. The error rate for German documents, however, is one to two to percent points higher. The error rate can be reduced by using several OCR results, e.g. by a voting algorithm.

The combination of several OCRs slows down the process, in particular for bad quality text areas.

Therefore this technique is only applied in small and important regions, e.g. date, addressee, or account number.

5 Focus of Attention

Important information in a document often is highlighted by typographic features in the text like font or style change, size, or by layout methods like use of white space, prominent position, or by graphical emphasis using lines and icons. The focus of attention (FOA) approach uses this knowledge to pre-sort the layout segments and to apply the OCR only to the relevant regions of interest. IDA presently uses FOA for extraction of address and date information. The information for guiding the FOA is available from the layout models, that are defined with the model acquisition tool (chapter 3).

6 Content Analysis

To cope with OCR errors we developed a fuzzy string matcher based on weighted error measures for substitution, insertion and deletion [1]. The new similarity measure considers the confusion probability matrix of OCR and the effect of entropy of characters in words [2].
Larger portions of text, e.g. the letter body, are analyzed by a new statistical method, which determines the relevance of words for a given category. This method can be applied to classify the textual information of the documents into categories without using keywords. The only prerequisite is a training with a sufficient number of samples for each category [2].

In the first processing step of a text we normalize the words with a lemmatizer. For each word the lemmatizer returns one or more stems and it's lexical category. Categorizing a text in our system depends on the relevance of words for categories. The relevance rlv(w in c) of a word w for a category c is defined by Pearsons well known correlation coefficient. The relevance of a word for various given categories is computed from a set of labeled training texts.

7 Applications and Results

The system has been applied to three different document types: forms, business letters and technical articles.
The forms application was part of a system solution for conversion of more than five million real estate register pages into an optical archive [8]. IDA solved the task to classify one of 10 different register page types, and to determine the location in each text column , where a new text entry can be made (Fig. 3). This allows fast access in long columns of the register, and the use of a hybrid editor, which shows the new text as overlay over the document image. The recognition speed per page is less than 0.6 sec. The task is finished and the error rate for the end marks is estimated to be lower than 0.1%.

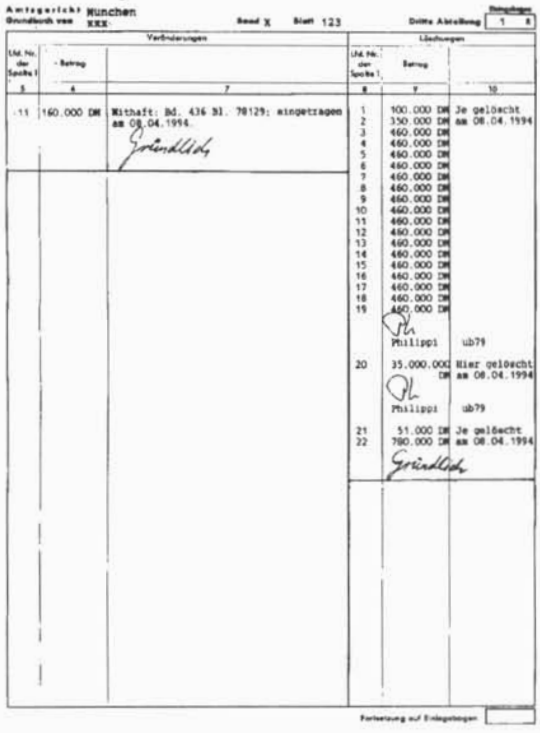

Figure 3: Example of real estate register page with recognized locations for continuation of text entry (indicated by rules below the signatures).

For business letters and articles IDA has solved the task to presort the scanned documents and to extract index information for archiving and retrieval applications (Fig. 1). The presorting was performed by layout classification, using models for business letters and title pages of articles from different journals. The index information was sender information and date for letters, and bibliographical data like title, author, etc. for articles. Presorting and getting sender information is very fast and accurate(about 0.3 sec., better 99%) using the layout model matcher with more than 200 models. Index information extraction has an error rate of about 3% and takes typical few seconds, depending on OCR. An interactive verification with prefilled index masks was necessary. Further classification of the letters into 16 application categories like invoice, order, etc. was performed by content analysis with an accuracy of 90% without rejection.
The latest application of IDA was in-house mail sorting of incoming business letters with the task to

distribute the letters to the destination department. For most letters IDA had to solve the following tasks: address block location on the first page of the letter and recognition of department name or alias if available. Using the layout analysis module and OCR in combination with the fuzzy string matcher, this was a state of the art problem. The test set consisted of 474 letters with 7 departments and 39 aliases. The department was determined with 0.6% error rate at 50% rejection.

There are, however, many letters without any department description in the address field. In these cases the content analysis of IDA was applied to the core message or body of the letter. The core message in letters is located between the salutation and the greeting/signature phrase. The OCR results of the letter bodies were labeled with the respective department by a mail distibution expert. The results are given in Table 3 together with other examples of text categories(number in parenthesis give the numbers of training-texts / test-texts / categories) where Recognition is the recognition rate without rejections, and Recall and Precision have the usual meanings. On a Sun Sparc-10 the categorizing algorithm has a performance of 1500 words/sec.

Application	Recog.	Recall	Precision
in-house mail (200/274/7)	59.5%	19.3%	97.9%
e-mail filing (221/296/9)	56.5%	32.1%	90.5%
dpa-news (600/450/4)	95.8%	82.5%	99.5%

Table 1: Text categorization results for some applications.

The results for mail distribution and e-mail filing mainly suffer from sparse training data. We are convinced that the retrieval performance of our text categorization can be significantly improved with larger training corpora and also with some extensions such as recognition of relevant multi-word terms and unambiguous word tagging for the lemmatizer.

8 Conclusions and Future Work

An overview on the architecture, applications and results of the document analysis system IDA has been given in this paper. The system has been implemented in C for Unix and PC platforms. The PC is preferred beause of the larger choice for available OCR engines. The overall error rates are about one to two percent, and the performance is about one second for each described task on a 100 MHz PC class computer. This time does not include OCR, which may add up to 20 sec per page, depending on the amount and quality of text.

The quality of character and word recognition for German has to be improved. The combination of different OCR results and the use of alternatives and confidence measures will be investigated further. This will increase the recognition rate of address interpretation for in-house mail sorting.

For the document classification methods using layout and text content, the lack of sufficient training samples has to be compensated by using more generic document knowledge.

The focus of attention approach will be extended to more highlighting techniques. For logos a subsystem for location and classification is in development.

Acknowledgment

Some topics of this work have been supported in part by the German Ministry of Education, Research and Technology (BMBF).

References

[1] A.A. Bertossi, F. Luccio, E. Lodi and L. Pagli, "String matching with weighted errors", Theoretical Computer Science 73, (1990), pp. 319-328.

[2] T. Brückner, P. Suda, H.U. Block, G. Maderlechner, "In-house Mail Distribution by Automatic Address and Content Interpretation", Proc. 5th. Ann. Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, USA, (1996), pp. 67 - 75.

[3] A. Dengel, et al., "OfficeMaid - A System for Office Mail Analysis, Interpretation and Delivery", Int. Workshop on Document Analysis Systems (DAS94), Kaiserslautern, Germany, (1994), pp. 253-275.

[4] S. Gopisetti, R. Lorie, J. Mao, M. Mohiuddin, A. Sorin, E. Yair, "Automated forms-processing software and services", IBM J. Res. Develop., Vol. 40, No. 2, March 1996, pp. 211 - 230.

[5] Annual Reports, Information Science Research Institute, University of Nevada, Las Vegas, 1993 - 1996.

[6] T. Pavlidis and J. Zhou, "Page Segmentation by White Streams", Proc. ICDAR91, Saint-Malo, France, (1991), pp. 945-950.

[7] J. Schürmann, "Text Recognition - From Pixels to Meaning", Proc. 5th. Ann. Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, USA, (1996), pp. 17 - 36.

[8] P. Suda, H. Bock, H.P. Klünder, G. Maderlechner, "How Can Document Analysis Help in Capturing Five Million Pages?", Proc. ICDAR95, Montreal, Canada, (1995), pp. 372 - 377.