Lecture Notes in Computer Science 2956

Edited by G. Goos, J. Hartmanis, and J. van Leeuwen

Springer
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Paris
Tokyo

Andreas Dengel Markus Junker
Anette Weisbecker (Eds.)

Reading and Learning

Adaptive Content Recognition

Springer

Series Editors

Gerhard Goos, Karlsruhe University, Germany
Juris Hartmanis, Cornell University, NY, USA
Jan van Leeuwen, Utrecht University, The Netherlands

Volume Editors

Andreas Dengel
Markus Junker
German Research Center for Artificial Intelligence
(DFKI GmbH)
P.O. Box 2080, 67608 Kaiserslautern, Germany
E-mail: {andreas.dengel,markus.junker}@dfki.de

Anette Weisbecker
University of Stuttgart
Institute for Human Factors and Technology Management
Fraunhofer Institute for Industrial Engineering
Nobelstr. 12, 70569 Stuttgart, Germany
E-mail: Anette.Weisbecker@iao.fhg.de

Library of Congress Control Number: 2004104165

CR Subject Classification (1998): I.5, I.4, H.3, H.4, I.2, I.7

ISSN 0302-9743
ISBN 3-540-21904-8 Springer-Verlag Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by PTP-Berlin, Protago-TeX-Production GmbH
Printed on acid-free paper SPIN: 10997871 06/3142 5 4 3 2 1 0

Foreword

The amounts of information that are flooding people both at the workplace and in private life have increased dramatically in the past ten years. The number of paper documents doubles every four years, and the amount of information stored on all data carriers every six years. New knowledge, however, increases at a considerably lower rate. Possibilities for automatic content recognition in various media and for the processing of documents are therefore becoming more important every day.

Especially in economic terms, the efficient handling of information, i.e., finding the right information at the right time, is an invaluable resource for any enterprise, but it is particularly important for small- and medium-sized enterprises. The market for document management systems, which in Europe had a volume of approximately 5 billion euros in 2000, will increase considerably over the next few years.

The BMBF recognized this development at an early stage. As early as in 1995, it pooled national capabilities in this field in order to support research on the automatic processing of information within the framework of a large collaborative project (READ) involving both industrial companies and research centres. Evaluation of the results led to the conclusion that research work had been successful, and, in a second phase, funding was provided for the collaborative follow-up project Adaptive READ from 1999 to 2003. The completion of these two important long-term research projects has contributed substantially to improving the possibilities of content recognition and processing of handwritten, printed and electronic documents.

Under the direction of the German Research Centre for Artificial Intelligence (DFKI) in Kaiserslautern, a total number of 15 partners, including 7 research centres and 4 SMEs, cooperated in the projects. The overall financial volume of both projects was 32 million euros, of which about 16 million euros were made available by the German Federal Ministry of Education and Research.

In an exemplary manner, the researchers successfully addressed both basic research and application-oriented development, and achieved major breakthroughs, which can be expected to promote the creation of high-tech jobs in the participating companies and to enhance Germany's attractiveness as a place for research and production. The projects have so far resulted in 110 scientific publications and in as many as 20 applications for patents or intellectual property rights. Owing to a considerable number of spin-off jobs, along with 20 spin-off products, including new reading devices for forms and business letters, the German companies participating in the project have become world leaders in this field.

The results of relevant research activities are of great importance in government provision for future needs (applications in the Federal Agency for Employment and in health insurance schemes), and moreover they offer new possibilities

for applications in industrial enterprises. More than half of all private health insurance companies that rely on electronic document processing are already using Adaptive READ systems. The research activities will also inspire and influence the development of new, intelligent technologies for Internet searching.

This report on the results of the research project is addressed to the scientific community with the aim of passing on the findings, and to business enterprises that intend to create high-tech jobs in Germany in this field.

December 2003

Dr. Bernd Reuse
Head of Software Systems Division
Federal Ministry of Education and Research

Preface

The information flood to be tackled every day by enterprises continues to increase. Every year about 200 million pages of paper worldwide are filed and more than 250 km of new folders are produced. The improvement of information management is considered as a relevant competitive factor by a growing number of enterprises. Enterprises demand a distribution of information in line with their needs and look for really relevant targeted statements and content. It is difficult to acquire relevant expertise, whether in intra-company networks, Internet use or traditional documentation, without efficient support systems that selectively cover information needs.

Existing interests, tasks and roles require that relevant information from multimedia documents of very different forms and structures has to be processed in such a way that it complies with individual requirements in the context of persons or enterprises. Furthermore, globalisation requires that information is available, independent of language, for very different missions and applications.

Under the heading "Reading and Learning – from Document to Knowledge", the research project Adaptive READ united German efforts from research and industry with the objective of elaborating comprehensive concepts for document opening systems that provide fitness for learning processes and putting them into practice in terms of prototypes. The improvements in identification techniques and component-based software development constitute additional focal points. Eleven partners, including four research organizations, worked together for three and a half years in Adaptive READ on this challenging task. This book provides a description of the trendsetting results of the Adaptive READ research project.

While in previous research projects the improvement and completion of reading abilities were the most important parts of the research, the goal in Adaptive READ had a much larger dimension. It was not the incremental improvement of existing algorithms for recording documents that was most important, but rather the formulation and prototype style of the implementation of concepts to make document analysis systems adaptive in practical scenarios. The motivation for such systems is obvious: Unlike previous systems, they are not subject to the "law of progressive mismatching" on the one hand, and on the other hand they need much less engineering capacity to sustain them. The goal of the project was to create and specify the basis for architectures, which, contrary to current document analysis systems with their predefined and frozen parameters and knowledge bases, are able to automatically adjust to the slowly changing characteristics of document layout, fonts, and writing habits. In systems in which automatic learning is not possible at all or only with difficulty, people have to take on this adaptation task. Adequate and intuitive user-friendly parameters, as well as significant analysis and diagnostics, should be provided to do this.

Unfortunately, however, a fully automated learning procedure was not feasible for a series of recognition tasks. In many cases (e.g., the need for random

sampling and its management) it was too costly. In these cases it was necessary to adjust the user-controlled instructions of the system so that they could be executed by a system administrator or end-user. Administration tools were needed for this, which make the recognition system cycle transparent. Also recognition methods were demanded, which, in the case of errors, make possible clear classifications according to the adaptation elements to which they belong (e.g., the parameters of a method). Moreover recognition systems must be easily tailored to the respective application areas.

Overview

After scanning, in order to be able to analyze documents electronically, they have to be aligned according to the orientation of the writing. In the first paper, "Error Tolerant Paper Deskew," Woitha and Janich present a solution to a particularly difficult case in document alignment – i.e., when the edges of the documents are mostly ruined. A further problem at a very early stage of document analysis is the differentiation between foreground and background information. In the second paper, "Adaptive Threshold," Woitha and Janich describe a promising technique that works with a dynamically defined threshold to distinguish between fore- and background information. Distinguishing between foreground and background in color documents is a special challenge. In "Neighborhood Related Color Segmentation Based on a Fuzzy Color Classification Tool" Frei, Hund and Schnitzlein describe the exploitation of specific background knowledge concerning the colored composition of documents and the document-specific color design for this task.

The contribution "Improving Image Processing Systems Using Artificial Neural Networks" by Rebmann, Michaelis, Krell, Seiffert and Püschel is concerned with two further problems involving the scanning of documents. A special type of hardware is introduced which, with the help of neural networks, shows ways of compensating for the effects of non-flat originals (e.g., books that are not fully opened). Furthermore, in this contribution they show, based on the techniques of neural networks, how to reduce picture artifacts as they emerge through the JPEG compression process.

When analyzing map material, the normal way of dividing fore- and background into two classes does not work. In "Adaptive Segmentation of Multicolored Documents Without a Marked Background" Eberhardt, Römer and Saedler show how information from color documents can be dissected into as many classes as required.

New techniques for handwriting recognition were also issues in the Adaptive READ project. In the paper "Recognition of Short Handwritten Texts," Boldt and Asp describe a new generation of handwriting readers that couple the old way of strict successive analysis steps, image processing, character recognition and syntactic alignment with feedback. The contribution "Handwritten Address Recognition Using Hidden Markov Models" by Brakensiek and Rigoll is devoted

to the problem of regionally different styles of writing for address recognition, and proposes various solutions using hidden Markov models.

Different systems for optical character recognition (OCR) usually have different strong points and weaknesses. The contribution "Adaptive Combination of Commercial OCR Systems" by Wilczok and Lellmann describes a flexible framework for the combination of recognition results of various OCR systems on the lexical level. It stands out particularly because words can be synchronized with the help of geometric criteria, which means that incorrect character segmentation can be avoided.

In the area of software engineering it appears that component-based approaches show promise in developing new application systems quickly and cheaply. In "Component-Based Software Engineering Methods for Systems in Document Recognition, Analysis and Understanding," Höß, Strauß and Weisbecker describe the bases of application of component-based approaches in document recognition systems. The design and implementation of a concrete and real component-based system for document recognition is described by Middendorf, Peust and Schacht in the contribution "A Component-Based Framework for Recognition Systems." The contribution "*smartFIX*: An Adaptive System for Document Analysis and Understanding" by Klein, Dengel and Fordan describes the smartFix system for extracting information from paper documents. It includes results on medical bills and prescriptions.

In "How Postal Address Readers Are Made Adaptive," Kreuzer, Miletzki, Schäfer, Schambach and Schulte-Austum deal with adaptive postal address readers. The ultimate goal in this area is to address readers, which, on the basis of reading the daily post, can adapt continuously to places, address structures, and handwriting. The paper also puts forward, with the support of Adaptive READ, solutions for cost reduction with the new services "forwarding" and "renumeration security."

The analysis of already existing electronic documents, such as e-mails or Web sites using learning methods was also of special interest in Adaptive READ. The contribution "A Tool for Semi-automatic Document Reengineering" by Drawehn, Altenhofen, Stanišić-Petrović and Weisbecker covers techniques for the semi-automatic structuring of full text documents. In "Inspecting Document Collections" Bohnacker, Franke, Mogg-Schneider and Renz describe further methods which help to analyze large document collections quickly, whereby documents that belong together are clustered and user-adaptive document-spanning combinations are generated. A model for so-called Collaborative Information Retrieval (CIR) was developed and resulted in a set of new search algorithms. The idea in CIR is to monitor search processes of a search engine in order to optimize future search processes – i.e., to find the desired information faster. Two different methods to find a solution to this problem are shown in the contributions "Introducing Query Expansion Methods for Collaborative Information Retrieval" by Hust and "Improving Document Transformation Techniques with Collaborative Learned Term-Based Concepts" by Klink. Whereas in these approaches the goal is to find relevant documents, the contribution "Passage Retrieval Based

on Density Distributions of Terms and Its Applications to Document Retrieval and Question Answering" by Kise, Junker, Dengel and Matsumoto focusses on finding relevant passages within documents.

Finally, in the framework of Adaptive READ two representative surveys were conducted in companies on tool application in the document management environment. Apart from other factors, the results described in "Results of a Survey About the Use of Tools in the Area of Document Management" by Altenhofen, Hofman, Kieninger and Stanišić-Petrović provide a profound insight into the infrastructural prerequisites in German companies, the current degree of utilization of various technologies, and the degree of user satisfaction.

December 2003

Prof. Dr. Andreas R. Dengel
Dr. Markus Junker
Priv.-Doz. Dr.-Ing. habil. Anette Weisbecker

Table of Contents