Handbook of Document Image Processing and Recognition

David Doermann • Karl Tombre

Editors

Handbook of Document Image Processing and Recognition

With 339 Figures and 98 Tables

Springer Reference

Editors
David Doermann   Karl Tombre
University of Maryland   Université de Lorraine
College Park, MD, USA   Nancy, France

Printed on acid-free paper

Foreword

In the beginning, there was only OCR. After some false starts, OCR became a competitive commercial enterprise in the 1950's. A decade later there were more than 50 manufacturers in the US alone. With the advent of microprocessors and inexpensive optical scanners, the price of OCR dropped from tens and hundreds of thousands of dollars to that of a bottle of wine. Software displaced the racks of electronics. By 1985 anybody could program and test their ideas on a PC, and then write a paper about it (and perhaps even patent it).

We know, however, very little about current commercial methods or in-house experimental results. Competitive industries have scarce motivation to publish (and their patents may only be part of their legal arsenal). The dearth of industrial authors in our publications is painfully obvious. Herbert Schantz's book, *The History of OCR*, was an exception: he traced the growth of REI, which was one of the major success stories of the 1960's and 1970's. He also told the story, widely mirrored in sundry wikis and treatises on OCR, of the previous fifty years' attempts to mechanize reading. Among other manufacturers of the period, IBM may have stood alone in publishing detailed (though often delayed) information about its products.

Of the 4000-8000 articles published since 1900 on character recognition (my estimate), at most a few hundred really bear on OCR (construed as *machinery - now software - that converts visible language to a searchable digital format*). The rest treat character recognition as a prototypical classification problem. It is, of course, researchers' universal familiarity with at least some script that turned character recognition into the pre-eminent vehicle for demonstrating and illustrating new ideas in pattern recognition. Even though some of us cannot tell an azalea from a begonia, a sharp sign from a clef, a loop from a tented arch, an erythrocyte from a leukocyte, or an alluvium from an anticline, all of us know how to read.

Until about 30 years ago, OCR meant recognizing mono-spaced OCR fonts and typewritten scripts one character at a time – eventually at the rate of several *thousand* characters per second. Word recognition followed for reading difficult-to-segment typeset matter. The value of language models more elaborate than letter n-gram frequencies and lexicons without word frequencies gradually became clear. Because more than half of the world population is polyglot, OCR too became multilingual (as Henry Baird predicted that it must). This triggered a movement to post all the cultural relics of the past on the Web. Much of the material awaiting conversion,

ancient and modern, stretches the limits of human readability. Like humans, OCR must take full advantage of syntax, style, context, and semantics.

Although many academic researchers are aware that OCR is much more than classification, they have yet to develop a viable, broad-range, end-to-end OCR system (but they may be getting close). A complete OCR system, with language and script recognition, colored print capability, column and line layout analysis, accurate character/word, numeric, symbol and punctuation recognition, language models, document-wide consistency, tuneability and adaptability, graphics subsystems, effectively embedded interactive error correction, and multiple output formats, is far more than the sum of its parts. Furthermore, specialized systems - for postal address reading, check reading, litigation, and bureaucratic forms processing - also require high throughput and different error-reject trade-offs. Real OCR simply isn't an appropriate PhD dissertation project.

I never know whether to call hand print recognition and handwriting recognition "OCR." but abhor *intelligent* as a qualifier for the latest wrinkle. No matter: they are here to stay until tracing glyphs with a stylus goes the way of the quill. Both human and machine legibility of manuscripts depend significantly on the motivation of the writer: a hand-printed income tax return requesting a refund is likely to be more legible than one reporting an underpayment. Immediate feedback, the main advantage of on-line recognition, is a powerful form of motivation. Humans still learn better than machines.

Document Image Analysis (DIA) is a superset of OCR, but many of its other popular subfields require OCR. Almost all line drawings contain text. An E-sized telephone company drawing, for instance, has about 3000 words and numbers (including revision notices). Music scores contain numerals and instructions like *pianissimo*. A map without place names and elevations would have limited use. Mathematical expressions abound in digits and alphabetic fragments like *log, limit, tan* or *argmin*. Good lettering used to be a prime job qualification for the draftsmen who drew the legacy drawings that we are now converting to CAD. Unfortunately, commercial OCR systems, tuned to paragraph-length segments of text, do poorly on the alphanumeric fragments typical of such applications. When Open Source OCR matures, it will provide a fine opportunity for customization to specialized applications that have not yet attracted heavy-weight developers. In the meantime, the conversion of documents containing a mix of text and line art has given rise to distinct sub-disciplines with their own conference sessions and workshops that target graphics techniques like vectorization and complex symbol configurations.

Another subfield of DIA investigates what to do with automatically or manually transcribed books, technical journals, magazines and newspapers. Although Information Retrieval (IR) is not generally considered part of DIA or vice-versa, the overlap between them includes "logical" document segmentation, extraction of tables of content, linking figures and illustrations to textual references, and word spotting. A recurring topic is assessing the effect of OCR errors on downstream applications. One factor that keeps the two disciplines apart is that IR experiments (e.g., TREC) typically involve orders of magnitude more documents than DIA

experiments because the number of characters in any collection is far smaller than the number of pixels.

Computer vision used to be easily distinguished from the image processing aspects of DIA by its emphasis on illumination and camera position. The border is blurring because even cellphone cameras now offer sufficient spatial resolution for document image capture at several hundred dpi as well as for legible text in large scene images. The correction of the contrast and geometric distortions in the resulting images goes well beyond what is required for scanned documents.

This collection suggests that we are still far from a unified theory of DIA or even OCR. The Handbook is all the more useful because we have no choice except to rely on heuristics or algorithms based on questionable assumptions. The most useful methods available to us were all *invented* rather than derived from prime principles. When the time is ripe, many alternative methods are invented to fill the same need. They all remain entrenched candidates for "best practice". This Handbook presents them fairly, but generally avoids picking winners and losers.

"Noise" appears to be the principal obstacle to better results. This is all the more irritating because many types of noise (e.g. skew, bleed-through, underscore) barely slow down human readers. We have not yet succeeded in characterizing and quantifying signal and noise to the extent that communications science has. Although OCR and DIA are prime examples of information transfer, information-theoretic concepts are seldom invoked. Are we moving in the right direction by accumulating empirical midstream comparisons – often on synthetic data – from contests organized by individual research groups in conjunction with our conferences?

Be that as it may, as one is getting increasingly forgetful it is reassuring to have most of the elusive information about one's favorite topics at arm's reach in a fat tome like this one. Much as on-line resources have improved over the past decade, I like to turn down the corner of the page and scribble a note in the margin. Younger folks, who prefer search-directed saccades to an old-fashioned linear presentation, may want the on-line version.

David Doermann and Karl Tombre were exceptionally well qualified to plan, select, solicit, and edit this compendium. Their contributions to DIA cover a broad swath and, as far as I know, they have never let the song of the sirens divert them from the muddy and winding channels of DIA. Their technical contributions are well referenced by the chapter authors and their voice is heard at the beginning of each section.

Dave is the co-founding-editor of *IJDAR*, which became our flagship journal when *PAMI* veered towards computer vision and machine learning. Along with the venerable *PR* and the high-speed, high-volume *PRL*, *IJDAR* has served us well with a mixture of special issues, surveys, experimental reports, and new theories. Even earlier, with the encouragement of Azriel Rosenfeld, Dave organized and directed the Language and Media Processing Laboratory, which has become a major resource of DIA data sets, code, bibliographies, and expertise.

Karl, another *IJDAR* co-founder, put Nancy on the map as one of the premier global centers of DIA research and development. Beginning with a sustained drive

to automate the conversion of legacy drawings to CAD formats (drawings for a bridge or a sewer line may have a lifetime of over a hundred years, and the plans for the still-flying Boeing 747 were drawn by hand), Karl brought together and expanded the horizons of University and INRIA researchers to form a critical mass of DIA.

Dave and Karl have also done more than their share to bring our research community together, find common terminology and data, create benchmarks, and advance the state of the art. These big patient men have long been a familiar sight at our conferences, always ready to resolve a conundrum, provide a missing piece of information, fill in for an absentee session chair or speaker, or introduce folks who should know each other.

The DIA community has every reason to be grateful to the editors and authors of this timely and comprehensive collection. Enjoy, and work hard to make a contribution to the next edition!

July 2013   George Nagy
                                                                   Professor Emeritus, RPI

Preface

Optical Character Recognition was one of the very first applications addressed in computer science, once this field extended beyond the needs for simulation and scientific computing, which boosted its birth during World War II. However, although real applications were quickly available, it progressively became clear to everybody that beyond the ability for a computer to decipher characters, there was a very vast domain of open scientific problems and hard technical challenges to be able to build complete document analysis systems, to extract the appropriate information from a vast range of documents meant to be converted from their original paper form to the appropriate digital formats.

As scientists, we have been very lucky to have been a part of this great scientific and technological adventure, throughout several decades. We have seen the field of document analysis and recognition emerge and mature, we have seen the community organize itself, and create conferences, workshops, and journals dedicated to this field. We have also seen companies integrate the best of the scientific results into great technological products, which have helped people and companies merge the paper world and the digital world, and invent new concepts of what a document is and how it is handled, stored, exchanged, and searched for.

In 2010, nearly 20 years had already passed since the first International Conference on Document Analysis and Recognition was held in Saint-Malo, France. We felt that it was time to consolidate the vast knowledge accumulated in the field into a comprehensive handbook making the state of the art available in a single volume.

It took us 3 years of work to put this book together, with the great help of a set of dedicated authors representing the very best experts in the field. The field of document image processing and recognition is primarily defined by the fact that the objects on which processing, recognition, and analysis methods are applied, are *documents* – usually captured by some kind of imaging process. But it relies on a vast set of knowledge about image processing, pattern recognition, data classification, and information retrieval. The temptation was often there to tell the reader more about the basics of these scientific areas. However, we came to the conclusion that this would lead us to putting together a complete encyclopedia, not just a handbook! Therefore, we assume that our readers are reasonably familiar with the fundamental methodology and concepts of these fields, or that they will look them up in other reference books.

We hope that the reader will find the final result as comprehensive and useful as we feel it is, thanks to the vast expertise gathered from all the senior experts who have contribute to this book.

July 2013   David Doermann
                                                                                        Karl Tombre

Acknowledgments

The journey that has now resulted in the first edition of the *Handbook of Document Image Processing and Recognition* began with a brief conversation with Wayne Wheeler almost 4 years ago. That meeting set the wheels turning on a project that has now involved dozens of people at all levels of the publication process. Certainly it would never have been possible to compile such a definitive collection of information, without the numerous experts that wrote the chapters. These researchers, some founders of the field, others the "next generation", are all volunteers, yet they made the time and put forth the effort necessary to establish this unparalleled reference. The community is indebted to them for their help with this project.

There were also a number of members of our community that provided input at the conceptual stage and others that provided very constructive feedback beyond their own chapters. While some of you, for various reasons, did not end up participating directly as authors, your input was very important in shaping the final product.

Most importantly, we would like to thank the production team at Springer. We thank Wayne Wheeler for convincing us to undertake this project and for his continuous motivation stating the need for such a resource. We thank associate editors Marion Kraemer and Barbara Wolf who provided much of the infrastructure support when the handbook was just becoming a reality and provided continued support and advice to educate us along the way. Finally, we would like to thank editorial assistant Saskia Ellis who coordinated the day-to-day interactions between the editors, the authors, and the production departments. Without that level of commitment and dedication on her part, we would neither have had the bandwidth nor the patience to complete this book.

<div align="right">

Coeditors
David Doermann
Karl Tombre

</div>

About the Editors

David Doermann is a member of the research faculty at the University of Maryland College Park. He received a B.Sc. degree in computer science and mathematics from Bloomsburg University in 1987, and a M.Sc. degree in 1989 in the Department of Computer Science at the University of Maryland, College Park. He continued his studies in the Computer Vision Laboratory, where he earned a Ph.D. in 1993. Since then, he has served as co-director of the Laboratory for Language and Media Processing in the University of Maryland's Institute for Advanced Computer Studies and as an adjunct member of the graduate faculty.

His team of researchers focuses on topics related to document image analysis and multimedia information processing. Their recent intelligent document image analysis projects include page decomposition, structural analysis and classification, page segmentation, logo recognition, document image compression, duplicate document image detection, image-based retrieval, character recognition, generation of synthetic OCR data, and signature verification. In video processing, projects have centered on the segmentation of compressed domain video sequences, structural representation and classification of video, detection of reformatted video sequences, and the performance evaluation of automated video analysis algorithms.

In 2002, Dr. Doermann received an Honorary Doctorate of Technology Sciences from the University of Oulu for his contributions to digital media processing and document analysis research. He is a founding co-editor of the *International Journal on Document Analysis and Recognition*, has the general chair or co-chair of over a half dozen international conferences and workshops, and is the general chair of the International Conference on Document Analysis and Recognition (ICDAR) to be held in Washington DC in 2013. He has over 30 journal publications and over 160 refereed conference papers.

Karl Tombre, born in Trondheim (Norway) in 1961, received a French *doctorat* (Ph.D. degree) in computer science in 1987. From 1987 to 1998, he was *chargé de recherche* (senior researcher) at INRIA, a French national institute devoted to research in computer science and applied mathematics. In September 1998, he joined *École des Mines de Nancy* at *Institut National Polytechnique de Lorraine* as full professor, and was head of its Computer Science Department from 2001 to 2007. From 2007 to 2012, he returned on secondment to INRIA, to hold the position of director of the Inria Nancy – Grand Est research center, a large research center (about 450 staff members) for computer science and applied mathematics. Since September 2012, he is vice-president of *Université de Lorraine*, one of the largest universities in France, resulting from the merger of four universities in the Lorraine region, and is in charge of economic partnerships and of international affairs.

Prof. Tombre's scientific expertise is in document image analysis and recognition, with a major focus on graphics recognition. In 2002, he founded a research group on graphics recognition and led this group until September 2007. He is also at the origin of a free software package for graphics recognition. In the period 2006–2008, he was the president of the International Association for Pattern Recognition (IAPR). Prof. Tombre is co-founder of the series of international workshops on graphics recognition, and co-chair of the two first workshops, in 1995 and 1997. A co-founder of the *International Journal on Document Analysis and Recognition* (Springer Verlag), he was one its editors-in-chiefs from its creation to August 2013. Karl Tombre is also member of several other editorial boards and has been invited to serve on numerous international conference committees.

Contents

Volume 2

Contributors

Henry S. Baird Computer Science and Engineering Department, Lehigh University, Bethlehem, PA, USA

Elisa H. Barney Smith Electrical & Computer Engineering Department, Boise State University, Boise, ID, USA

Abdel Belaïd Université de Lorraine – LORIA, Vandœuvre-lès-Nancy, France

Dorothea Blostein School of Computing, Queen's University, Kingston, Canada

Horst Bunke Research Group on Computer Vision and Artificial Intelligence (FKI), Institute of Computer Science and Applied Mathematics, University of Bern, Bern, Switzerland

Huaigu Cao Raytheon BBN Technologies, Cambridge, MA, USA

Bertrand Coüasnon IRISA/INSA de Rennes, Rennes Cedex, France

Niladri Sekhar Dash Linguistic Research Unit, Indian Statistical Institute, Kolkata, India

Andreas Dengel German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, Germany

David Doermann Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA

Haikal El Abed Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany

Alicia Fornés Computer Vision Center & Computer Science Department, Universitat Autònoma de Barcelona, Bellaterra, Spain

Volkmar Frinken Computer Vision Center, Autonomous University of Barcelona, Bellaterra, Spain

Basilis G. Gatos Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", Agia Paraskevi, Athens, Greece

Michel Gilloux Seres/Docapost EBS (Groupe La Poste), Nantes, France

Venu Govindaraju Department of Computer Science & Engineering, Center for Unified Biometrics and Sensors, University at Buffalo, Amherst, NY, USA

Jianying Hu IBM T.J. Watson Research Center, Yorktown Heights, NY, USA

Donato Impedovo Politecnico di Bari, Bari, Italy

Dimosthenis Karatzas Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra, Spain

Anastasios Kesidis Department of Surveying Engineering, Technological Educational Institution of Athens, Aigaleo, Athens, Greece

Jin Hyung Kim Department of Computer Science, Korea Advanced Institute of Science and Technology, Yuseong-gu, Daejeon, Korea

Koichi Kise Department of Computer Science and Intelligent Systems, Graduate School of Engineering, Osaka Prefecture University, Sakai, Osaka, Japan

Bart Lamiroy Université de Lorraine – LORIA, Vandœuvre-lès-Nancy Cedex, France

Aurélie Lemaitre IRISA/Université Rennes 2, Rennes Cedex, France

Linlin Li National University of Singapore, Singapore, Singapore

Ying Liu Korea Advanced Institute of Science and Technology (KAIST), Yuseong-gu, Daejeon, Republic of Korea

Josep Lladós Computer Vision Center & Computer Science Department, Universitat Autònoma de Barcelona, Bellaterra, Spain

Tong Lu Department of Computer Science and Technology, State Key Laboratory for Novel Software Technology at Nanjing University, Nanjing, China

Volker Märgner Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany

Simone Marinai Dipartimento di Ingegneria dell'Informazione, Università degli studi di Firenze, Firenze, Italy

Prem Natarajan Information Sciences Institute, University of Southern California, Marina Del Rey, CA, USA

Nicola Nobile Centre for Pattern Recognition and Machine Intelligence (CENPARMI), Concordia University, Montréal, QC, Canada

Jean-Marc Ogier Université de La Rochelle, La Rochelle Cédex 1, France

Umapada Pal Computer Vision And Pattern Recognition Unit, Indian Statistical Institute, Kolkata, India

Giuseppe Pirlo Dipartimento di Informatica, Università degli Studi di Bari, Bari, Italy

Réjean Plamondon Département de génie électrique, Ecole Polytechnique de Montréal, Montréal, QC, Canada

Oriol Ramos Terrades Universitat Autònoma de Barcelona – Computer Vision Center, Bellaterra, Spain

Mohamed Imran Razzak Department of Computer Science and Engineering, Air University, Pakistan

College of Public Health and Health Informatics, King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia

Marçal Rusiñol Computer Vision Center & Computer Science Department, Universitat Autònoma de Barcelona, Bellaterra, Spain

Gemma Sánchez Computer Vision Center & Computer Science Department, Universitat Autònoma de Barcelona, Bellaterra, Spain

Srirangaraj Setlur Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Buffalo, NY, USA

Faisal Shafait Multimedia Analysis and Data Mining Competence Center, German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, Germany

School of Computer Science and Software Engineering, The University of Western Australia, Perth, Australia

Zhixin Shi Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Buffalo, NY, USA

Bong-Kee Sin Department of IT Convergence and Applications Engineering, Pukyong National University, Namgu, Busan, Korea

Ching Y. Suen Department of Computer Science and Software Engineering, Centre for Pattern Recognition and Machine Intelligence (CENPARMI), Concordia University, Montréal, QC, Canada

Salvatore Tabbone Université de Lorraine – LORIA, Vandœuvre-lès-Nancy Cedex, France

Chew Lim Tan Department of Computer Science, National University of Singapore, Singapore, Singapore

Karl Tombre Université de Lorraine, Nancy, France

Sergey Tulyakov Center for Unified Biometrics and Sensors, University at Buffalo, Amherst, NY, USA

Seiichi Uchida Department of Advanced Information Technology, Kyushu University, Nishi-ku, Fukuoka, Japan

Ernest Valveny Departament de Ciències de la Computació, Computer Vision Center, Universitat Autònoma de Barcelona, Bellaterra, Spain

Liu Wenyin College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai, China

Richard Zanibbi Department of Computer Science, Rochester Institute of Technology, Rochester, NY, USA

Xi Zhang National University of Singapore, Singapore, Singapore