Digital Libraries and Document Image Analysis

Henry S. Baird
Computer Science & Engineering Dept., Lehigh University
Bethlehem, Pennsylvania, USA

Abstract

The rapid growth of digital libraries (DLs) worldwide poses many challenges for document image analysis (DIA) research and development. DLs promise to offer more people access to larger document collections, and at far greater speed, than physical libraries can. But DLs also tend, for many reasons, to serve poorly documents which, although readily legible to people, are not accurately digitally encoded. Originally printed and handwritten documents, for example, in their original physical (undigitized) form of ink-on-paper are widely preferred, over electronic displays, for reading and other uses, whereas in the form of document images accessed through DLs they lose many of these advantages while of course lacking advantages of 'born digital' documents. This talk explores these issues and illustrates them with case studies arising from the author's experience as a DIA researcher in collaboration with several DL projects in the US. The pace and scale of commercial document-scanning projects has been accelerating over the last three years. Difficult open DIA technical problems in DL applications are identified in the contrasting advantages of paper and digital displays, at every stage of capture, early processing, recognition, analysis, presentation, & retrieval, and in personal and interactive applications. Discussions at Int'l Workshop on Document Image Analysis for Libraries (DIAL2004), recently organized by Prof. Venu Govindaraju and the author, are summarized.

1. Open Problems in Document Image Analysis for DLs

Serious technical obstacles prevent *imaged* paper documents from playing all of the useful roles in digital libraries (DLs) that symbolically *encoded* documents can, or even many of the roles that are easy and natural for paper itself. We discuss the most difficult and urgent open problems for document image analysis (DIA) R&D applicable to DLs. Fundamental research is needed into the relative advantages of physical (non-digital) document media compared to encoded digital media. The conditions of document-image capture have consequences for human and machine legibility, completeness of collections, support for scholarly study, and archival conservation. Many challenging problems arise in early image processing in support of quality control and compression. There are a large number of obstacles to the fully automatic, high-accuracy analysis of the content of document images in DL contexts. Improved methods for presentation, display, printing, and reflowing of document images are needed. High-performance retrieval, indexing, and summarization of document–images challenges rhe state of the art of DIA technology. Special problems arise in 'personal' and interactive digital libraries.

Many physical properties of ink-on-paper assist human reading,[8] *e.g.* lightweight, thin, flexible, markable, unpowered (and so 'always-on'), stable, cheap, etc. Of course the digital display devices used to access today's DLs — desktop, laptop, and handheld computers, plus eBook readers, tablet PCs, etc — have many advantages too: automatically and rapidly rewritable, interactive, connected (*e.g.* wirelessly) via networks to vast databases, etc. The many ways in which information conveyed originally as ink-on-paper may, and *may not*, be better delivered by digital means need to be better elucidated (for an extended discussion, see Ref. [3]). Still, it is by no means certain[8] that any digital delivery of document images can compete with paper.

The capture of document images for use in DLs often occurs in large-scale batch operations. For reasons of cost, only rarely will the documents be rescanned. In fact, documents can be damaged or destroyed in the process, sometimes deliberately. It is thus urgent to design document scanning operations so that the resulting images will serve a wide variety of uses for many years, not merely those uses most immediately in mind at the time. Image quality is most often quantified through the technical specifications of the scanning equipment, *e.g.* depth/color, color gamut and calibration, lighting conditions, digitizing resolution, compression method, and image file format. Such measures are vitally important but they are the means of quality control, not the end. Research is needed into *goal-directed metrics* of document image quality, tied quantitatively to the reliability of downstream processing (both machine and human) of the images.

The technical specifications of scanning conditions should be preserved and attached (as metadata) to the resulting images. Tools for the automatic estimation of scanner parameters from images of text could be an important contribution. Exploratory research in this direction is under way (*e.g.* Ref. [9]), and a few DIA studies have attempted to predict OCR performance and to choose image restoration methods to improve OCR, guided by automatic analysis of images (cf. Ref. [10] and its references). The results so far are not negative, but the gains are modest. Can these methods be refined to produce large improvements? Can

improving image quality, by itself, improve OCR results enough to obviate the need for post-OCR correction?

A wide range of early-stage image processing tools are needed to support high-quality image capture. Image calibration and restoration must usually be specialized to the scanner. Image processing should, ideally, occur quickly enough for the operator to check each page image visually for consistent quality. Tools are needed for orienting pages so text is rightside up, deskewing the page, removing some of the pepper noise, and removing dark artifacts on or near the image edges, and so forth.

The analysis and recognition of the content of document images requires, of course, the full range of DIA R&D achievements: page layout analysis, text/non-text separation, printed/handwritten separation, text recognition, labeling of text blocks by function, automatic indexing and linking, table and graphics recognition, etc. Most of the DIA literature is devoted to these topics.

The central task of DIA research has long been to extract a full and perfect transcription of the textual content of document images. No existing OCR technology, experimental or commercially available, can guarantee near-perfect accuracy across the full range of document images of interest to users. It is rarely possible to predict how badly an OCR system will fail on a given document. Even worse, it is usually impossible to estimate automatically, after the fact, how badly an OCR system has performed (but, see [7]). This combination of unreliability, unpredictability, and untrustworthiness requires expensive manual 'proofing' (inspection and correction) in document scan-and-conversion projects that require a uniformly high standard of accuracy. The open problems here are clearly difficult, urgent, and many, but they are also thoroughly discussed in the DIA literature (*e.g.* Refs. [6] and [5]).

Determining the reading order among blocks of text is of course a DIA capability critically important for DLs since it would allow more fully automatic navigation through images of text. This however remains an open problem in general, in that a significant residue of cases cannot be disambiguated through physical layout analysis alone, but seem to require linguistic or even semantic analysis.

Detecting and analyzing tabular data is a problem which has received sustained attention by the DIA community. It is of course harder in general than the analysis of images of body text; it appears however to be far easier than detecting and analyzing images of mathematical notation.

In the most general case, DLs would benefit from DIA facilities that label every part of document structure within images to a degree of refinement possible using markup languages such as XML — this remains a resistant class of DIA problems.

Recently, DIA researchers have investigated systems for the automatic analysis of document images into image fragments (*e.g.* word images) that can be reconstructed or "reflowed" onto a display device of arbitrary size, depth, and aspect ratio (*e.g.* Ref. [2]). It would be highly useful to extend reflowing to other parts of document images, such as tables and graphics, difficult as it may be to imagine, at the present state of the art, how this could be accomplished.

The indexing and retrieval of document images are critical for the success of DLs. Most published methods (surveyed in Ref. [4]) for retrieval of document images first attempt recognition and transcription followed by indexing and search operating on the resulting (in general, erroneous) encoded text (using, *e.g.*, standard 'bag-of-words' information retrieval (IR) methods). An open problem, not much studied, is the effectiveness of OCR–¿IR methods on short passages, such as, in an extreme but practically important case, fields containing key metadata (such as title, author, etc).

Research has recently gotten underway in 'personal digital libraries,' with the aim of offering tools to individuals willing to try to scan their own documents and, mingling imaged and encoded files, assemble and manage their own DLs.

As publically available DLs gather large collections of document images, opportunities will arise for collective improvement of the DL services. Within such a community of volunteers, assuming it could establish a culture of trust, review, and acceptance, DIA tools could be critically enabling. To assist such interactive projects, the DIA field should consider developing DIA tool sets freely downloadable from the web, or perhaps run on DL servers on demand from users. In this way even very large collections of document images could be improved beyond the level possible today through exclusively automatic DIA processing.

2. The DIAL2004 Workshop

The first International Workshop on Document Image Analysis for Libraries (DIAL2004, January 23-24, 2004, Palo Alto, CA) brought together fifty-five researchers, end-users, practitioners, businessmen, and end-users who were all interested in new technologies assisting the integration of imaged documents within DLs so that, ideally, everything that can be done with 'born digital' data can also be done with scanned hardcopy documents. Academia, industry, and government in twelve countries were represented by researchers from the document image analysis, digital libraries, library science, information retrieval, data mining, and humanities fields. The participants worked together, in panels, debates, and group discussions, to describe the state of the art and identify urgent open problems. More broadly, the workshop attempted to stimulate closer cooperation in future between the DIA and DL communities.

Twenty-nine regular papers, published in the proceedings,[1] established the framework of discussion, which embraced six broad topics:

- DIA Challenges in Historical DL Collections;
- DIA Challenges in DLs of Handwritten Documents; and
- Multilingual DLs.
- DIA Challenges within DLs;
- DL Systems Architectures & Costs;
- Retrieval in DLs using DIA Methods;
- Content Extraction from Document Images for DLs;

The remaining sections of this paper summarize work relating to these topics with special emphasis on discussions that took place at DIAL2004 on the first three topics.

3. Acknowledgment

This paper attempts to synthesize knowledge and advice received from from many people at PARC, the University of California, Berkeley, Digital Libraries Initiative group, the Carnegie–Mellon University DL community, the participants in the Working Group on Digital Libraries, held at the IAPR Workshop on Document Analysis Systems, Princeton, NJ, August, 2002, and many others.

References

1. H. S. Baird and V. Govindaraju, editors. *Proceedings, 1st Int'l Workshop on Document Image Analysis for Libraries (DIAL2004)*. IEEE Computer Society Press, Palo Alto, CA, January 2004.
2. T. M. Breuel, W. C. Janssen, K. Popat, and H. S. Baird. Paper to PDA. In *Proc., IAPR 16th ICPR*, pages Vol. 4, 476–479, Quebec City, Canada, August 2002.
3. A. Dillon. Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics*, 35(10):1297–1326, 1992.
4. D. Doermann. The indexing and retrieval of document images: A survey. *Computer Vision and Image Understanding*, 70(3), June 1998. Special Issue on "Document Image Understanding and Retrieval," J. Kanai and H. S. Baird (Eds.).
5. G. Nagy. Twenty years of Document Image Analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
6. S. V. Rice, F. R. Jenkins, and T. A. Nartker. The fifth annual test of OCR accuracy. Technical report, Information Science Research Institute, Univ. of Nevada at Las Vegas, Las Vegas, Nevada, 1996. ISRI TR-96-01.
7. P. Sarkar, H. S. Baird, and J. Henderson. Triage of ocr output using 'confidence' scores. In *Proc., 9th IS&T/SPIE Document Recognition & Retrieval Conf.*, San Jose, CA, January 2002.
8. A. J. Sellen and R. H. R. Harper. *The Myth of the Paperless Office*. The MIT Press, Cambridge, MA, 2002.
9. E. H. B. Smith and X. Qiu. Relating statistical image differences and degradation features. In *Proceedings, 5th IAPR International Workshop on Document Analysis Systems*, pages 1–12, Princeton, NJ, August 2002. Springer Verlag. LNCS 2423.
10. K. Summers. Document image improvement for OCR as a classification problem. In T. Kanungo, E. H. B. Smith, J. Hu, and P. B. Kantor, editors, *Proc., SPIE/IS&T Electronic Imaging Conf. on Document Recognition & Retrieval X*, pages 73–83, Santa Clara, CA, January 2003. SPIE Vol. 5010.

baird@cse.lehigh.edu