Document Analysis Systems Architectures for Digital Libraries

Henry S. Baird[1], Venugopal Govindaraju[2], and Daniel P. Lopresti[1]

[1] CSE Department, Lehigh University, Bethlehem, PA, USA
[baird,lopresti]@cse.lehigh.edu
[2] CEDAR, University at Buffalo, SUNY, Buffalo, NY, USA
govind@cedar.buffalo.edu

Abstract. Implications of technical demands made within digital libraries (DLs) for the architecture of document image analysis systems are discussed. The state-of-the-art is summarized, including a digest of themes that emerged during the recent International Workshop on Document Image Analysis for Libraries (DIAL2004). We attempt to specify, in considerable detail, the essential features of document analysis systems that can assist in: (a) the creation of DLs; (b) automatic indexing and retrieval of doc-images within DLs; (c) the presentation of doc-images to DL users; (d) navigation within and among doc-images in DLs; and (e) effective use of personal and interactive DLs.

1 Introduction

Within digital libraries (DLs) *imaged* paper documents are growing in number and importance, but they are too often unable to play many of the useful roles that symbolically *encoded* ("born digital") documents do. Traditional document image analysis (DIA) systems can relieve some but not all of these obstacles. In particular, the unusually wide variety of document images found in DLs, representing many languages, historical periods, and scanning regimes, taken together pose an almost insuperable problem for present-day DIA systems. How should DIA systems be redesigned to assist in the solution of a far broader range of DIA problems than have ever been attempted before?

Section 2 summarizes the principal points, relevant to this question, that were aired at the International Workshop on Document Image Analysis for Libraries (DIAL2004). The issue of hardcopy books versus digital displays is raised in Section 3. Section 4 considers problems associated with document-image capture, legibility, completeness checking, support for scholarly study, and archival conservation. Certain problems arising in early image processing may require fresh DIA architectures, described in Section 5. Section 6 points out implications for DIA systems of the lack of fully automatic, high-accuracy methods for analyzing doc-image content. Needs for improved methods for presentation, display, printing, and reflowing of document images are discussed in Section 7. Retrieval, indexing, and summarization of doc-images is addressed in Section 8. Section 9 lists problems arising in "personal" and interactive digital libraries.'

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, business people, 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 the future between the DIA and DL communities.

Twenty-nine regular papers, published in the proceedings [7], established the framework of discussion, which embraced six broad topics:

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

The remainder of this paper summarizes work relating to these topics, with special emphasis on discussions that took place at DIAL2004 on the first three.

2.1 Computational Issues in Historical DLs

Image Acquisition Image capture from historical artifacts needs special handling to counter the defects of document aging, and the physical constraints of digitization. A DIA oriented approach is suggested to effectively increase resolution, digitization speed, ensure document preservation (during scanning) and quality control [6, 35].

Bourgeouis *et al.* [35] use Signal to Noise Ratio (SNR) and other measures to demonstrate the loss of resolution/data in image compression formats, and recommend storage in 256 gray levels or true color. They observe that curators should be informed about the needs of DL technology and drawbacks of lossy file formats like JPEG. In addition, non-UV cold lights and automatic page turners are used to safeguard originals during scanning, and errors are countered by using skew, lighting and curvature correction for book bindings and color depth reduction for medieval documents. Character reconstruction is suggested to restore broken characters in ancient documents.

Continuous scanning is followed by automatic frame cropping as an efficient and fast procedure to generate images from microfilm [9]. Fourier-Mellin transform is used to correct rotation/shear, scale and translation errors [28]. Morphological operations, analysis of lightness and saturation in HLS (Hue, Lightness, and Saturation) image data, and connected component analysis is used to remove reconstructed paper areas [5].

Layout Analysis and Meta-Data Extraction Layout analysis and meta-data extraction is a crucial step in creating an information base for historical DL's. Even as researchers are gaining ground on complete recognition of text content from historical documents (Subsection 2.2), practical systems have been built using only the layout analysis stage of DIA [9, 26, 35].

Availability of images makes it possible to provide content based image retrieval, using even structural features like color and layout. Marinai *et al.* [39] create an MXY tree structure during document segmentation and then use layout similarity as a feature to query documents by example.

A historical DL should supplement content with meta-data describing textual features (*e.g.*, date, author, place) and geometrical information (*e.g.*, paragraph locations, image zones). Couasnon *et al.* use an automated web-based system for collecting annotations of French archives [18]. The system combines automatic layout analysis with human-assisted annotation in a web interface.

Transcription of historical documents maps ASCII text to corresponding words in the document image. This is intended to circumvent the lack of perfect Optical Character Recognition (OCR) for ancient writing styles [23, 33, 66].

2.2 Handwriting Recognition for DLs

Although commercial recognition products are available for printed script, handwriting recognition has achieved commercial success only in specialized domains. HMM based character model recognizers are used in address recognition from mail-piece images [51, 57]. This system relies on context information related to addresses.

For transcript creation from historical documents, mapping systems use handwriting recognition. OCR engines used in these applications cannot meet real-time recognition requests. Automatic author classification systems [65] use multi-stage binarization followed by identification of document writers using character features. For Hanja scripts, OCR and UI techniques [31] incorporate nonlinear shape normalization, contour direction features and recognizers based on Mahanalobis distance to generate transcripts for Hanja (Korean) documents.

A HMM based recognizer for large lexicons is examined for indexing historical documents in [23]. The system uses substring sharing, where a prefix tree is built from the lexicon. Entries that share the same prefix also share its computation without invoking the recognizer. Duration constraints on character states, choice pruning, and parallel decoding provide a speedup of 7.7 times.

Zhang *et al.* [66] combine word model recognition and transcript mapping to create handwritten databases. Lavrenko *et al.* [34] suggest a holistic recognition technique wherein normalized word images are used as inputs to a HMM. Scalar and profile features are extracted from the images and an entire historical document is modeled as a HMM, with words constituting the state sequence. For a document written by a given author, state transition probabilities are obtained by averaging word bi-gram probabilities collected from contemporary texts and previously transcribed writings of the target author.

2.3 Multilingual DLs

Despite excellent advances in Latin script DLs, research in other scripts such as Indic (Arabic, Bengali, Devanagari, and Telugu), Chinese, Korean, etc. is only recently receiving attention. Digital access to documents in these scripts is challenging by way of user interface (UI) design, layout analysis, and OCR.

A multi-lingual DL system should support simultaneous storage, entry, and display of data in many scripts. Many non-Latin scripts have a complicated character set [17], and need a separate encoding system. The display and entry of these languages requires new fonts [40, 47] and character input schemes. Also, to ensure compatibility and platform independence of data, a DL should not resort to customized solutions without completely examining existing standards.

In terms of character encoding, the Unicode[3] consortium aims at providing a reliable encoding scheme for all scripts in the world. It currently supports all commercial scripts and is accepted as a system standard by many DL researchers and software manufacturers [11, 32, 36, 40, 60, 63]. Although alternate schemes have been suggested [43], they do not have the compatibility and global acceptance of Unicode. On the storage front, XML is emerging as a versatile and preferred scheme for DL projects [3, 32, 53, 63].

Turning to input and display techniques, multi-layered input schemes for phonetic scripts [52] are suggested for stylus/keypad based entry systems (*e.g.*, for PDA's). Keyboard mapping systems (INSCRIPT for Indic scripts) map the keys of a standard $QWERTY$ keyboard onto the characters of a target script [43]. This keyboard system is functional, but has a steep learning curve. Moreover, every keyboard has to be physically labeled before a user can associate the keys with relevant characters. TrueViz [36] uses a graphical keyboard for Russian script input. Kompalli *et al.* [32] use a transliteration scheme, where Devanagari characters are entered by phonetic equivalent strings in English. For example, the Devanagari character क is entered using the English equivalent ka. A GUI keyboard is also provided to enter special characters.

The ability to display multiple languages on a single interface is dependent on the encoding schema and fonts used in a DL system. Most designers of multi-lingual software resort to Unicode based fonts, and software vendors provide detailed guidelines for internationalization [24].

2.4 Layout Analysis

Variation in the writing order of scripts, and the presence of language-specific constructs such as shirorekha (Devanagari), modifiers (Arabic and Devanagari), or non-regular word spacing (Arabic and Chinese) require different approaches to

[3] Unicode is an official implementation of ISO/IEC 10646.

layout analysis. For instance, gaps may not be used to identify words in Chinese and Arabic. Techniques for script identification vary from identifying scripts of individual words in a multi-lingual document [42] to those that determine scripts of lines [44] and entire text blocks [27, 62]. Once a script is identified, script-specific line and word separation algorithms can be used [22].

2.5 Multilingual OCR

Creation of data sets [30, 32] is a welcome development in providing training and testing resources for non-Latin script OCR. Providing data sets for certain scripts is a non-trivial task due to their large character sets and the variety of recognition units used by researchers [8, 13, 14, 38]. Some suggest splitting ground truth into components to provide truth at multiple levels of granularities [22].

Common methods for Indic script OCR use structural features to build decision trees [13, 14] or combine multiple knowledge bases to create statistical classifiers [8, 38]. Govindaraju *et al.* [22] combine structural and statistical features in a hybrid recognizer. Character images are pre-classified into categories based on structural features. A three layer Neural Network or a Nearest Neighbor classifier is then used to recognize the images.

Partial character matching is used for Chinese OCR [64]. When a character is presented to the recognizer, radicals or parts of characters are first identified. Classification of a sufficiently large number of components leads to recognition of the whole character. Tai *et al.* [61] use a multilayer perceptron network to divide Chinese characters into four layers. Classification at the lowest levels is followed by logical reconstruction to recognize characters.

Holistic techniques are being used for off-line and online recognition of Arabic [1, 4]. Psuedo 2D HMM's are used for ligature modeling in online recognition of Hangul scripts [54]. Bazzi *et al.* [10] recognize Arabic and English, using word-based HMM's with trigram character probabilities to improve recognition rates.

3 Ink-on-Paper *versus* Digital Displays

Many physical properties of ink-on-paper assist human reading [50], *e.g.*, lightweight, thin, flexible, markable, unpowered (and so "always-on"), stable, and cheap. Of course, digital display devices used to access today's DLs – desktop, laptop, and handheld computers, plus eBook readers, tablet PCs, etc. – have many advantages, too: they are automatically and rapidly rewritable, interactive, and connected (*e.g.*, wirelessly) via networks to vast databases. However, there remain many ways in which information conveyed originally as ink-on-paper *may not* be better delivered by digital means: these need to be better elucidated (for an extended discussion, see [19]).

It is by no means certain that any digital delivery of document images can compete with paper for all, or even for the most frequent purposes. It is still true today, as Sellen and Harper [50] report, that "paper [remains] the medium of choice for reading, even when the most high-tech technologies are to hand." They suggest these reasons:

1. paper allows “flexible [navigation] through documents;”
2. paper assists “cross-referencing” of several documents at one time;
3. paper invites annotation; and
4. paper allows the “interweaving of reading and writing.”

New technologies such as E-ink [21] and Gyricon [25] promise electronic document display with more of the advantages of paper (and new advantages of electronics). More – perhaps even fundamental – research into user-interactions with displays during reading and browsing appears to be needed to understand fully the obstacles to the delivery of document images via DLs.

4 Capture

Since the capture of document images for use in DLs usually occurs in large-scale batch operations during which documents may be damaged or destroyed, and which are too costly ever to be repeated, there is a compelling need for methods of designing document scanning operations so that the resulting images will serve a wide variety of uses for many years, not just those uses narrowly imagined at the time. Image quality should be – but often is not – carefully quantified, *e.g.*, at a minimum: depth/color, color gamut and calibration, lighting conditions, digitizing resolution, compression method, and image file format. In addition to these, we need richer *use-specific metrics* of document image quality, tied quantitatively to the reliability of downstream uses (*e.g.*, legibility, both machine and human).

4.1 Scanner Specifications

Digitizing resolutions (spatial sampling frequency) for textual documents typically range today between 300 and 400 pixels/inch (ppi); 600 ppi is less common but is gaining as scanner speed and disk storage capacity increase.

For what downstream uses are these rough guidelines sufficient? Research opportunities here are many, of this general type: does a particular scanning regime for modern books and printed documents (*e.g.*, 300 ppi 24-bit color) reliably provide images (of text, at least) which will support the best achievable recognition accuracy in the future, as image processing methods improve? Or should we, as a research community, help develop more exacting scanning standards?

A joint activity between AIIM and the Association for Suppliers of Printing, Publishing and Converting Technologies (NPES) is discussing an international standard (PDF-Archive) [45] to define the use of PDF for archiving and preserving documents.

Test targets for evaluating scanners include:

– IEEE Std 167A-1987, a facsimile machine test target that is produced by continuous-tone photography, with patterns and marks for a large range of measurements of moderate accuracy;

– AIIM Scanner Target, an ink-on-paper, halftone-printed target; and
– RIT Process Ink Gamut Chart, a four-color (cyan, magenta, yellow, and black), halftone-printed chart for low accuracy color sensitivity determinations.

To what extent do existing test targets, *e.g.*, AIIM [2] ANSI/AIIM MS-44-1988 "Recommended Practice for Quality Control of Image Scanners" and MS-44, allow for the manual or automatic monitoring of image quality needed for DIA processing? Do we need to design new targets for this purpose?

4.2 Measurement and Monitoring of Quality

Certainly we must recommend that the technical specifications of scanning conditions be preserved and attached (as metadata) to the resulting images. For many existing databases of document images, this has not been done. To our knowledge there does not yet exist a recommendation for such standards. Therefore, tools for the automatic estimation of scanner parameters from images of text could be an important contribution to the success of DLs. Exploratory research in this direction is under way (*e.g.*, [55]), but many questions are as yet unanswered, for example, how accurate will these estimates be? Can we estimate most of the image quality parameters that affect recognition? Will they run fast enough to be applied in real time as the images are scanned?

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. [59] and its references). The gains, so far, 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?

5 Initial Processing

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, and sometimes to the batch. Image processing should, ideally, occur quickly enough for the operator to check each page image visually for consistent quality; this modest capability is, as of yet, hard to achieve. Tools are needed for orienting pages so text is rightside up, for deskewing, deshearing, and dewarping, and for removing pepper noise and dark artifacts in book gutters and near edges of images. Software support for clerical functions such as page numbering and ordering, and the collection of metadata, are also crucial to maintaining high throughput. Few, if any, of these tasks present difficult DIA problems, but care is needed in the design of the user interface.

One place where DIA technology could help is in checking each page image for completeness and consistency: (a) Has any text been unintentionally cropped? (b) Are basic measures of image consistency (*e.g.*, brightness, contrast, intensity histograms) stable from page to page, hour after hour? (c) Are image properties consistent across the full page area for each image? These seem to be fairly challenging problems in general, but specific cases may yield to standard image processing techniques.

Are the page numbers – located and read by OCR on the fly – in an unbroken ascending sequences, and do they correspond to the automatically generated metadata? This problem is surely directly solvable using existing techniques, with perhaps the addition of string-correcting constraint-satisfaction analysis of the number sequences: however, we are not aware of any published solution. Perhaps it will someday be possible to assess both human and machine legibility on the fly (today this may seem a remote possibility, but cf. [16]).

5.1 Restoration

Document image restoration can assist fast and painless reading, OCR for textual content, DIA for improved human reading (*e.g.*, format preservation), and characterization of the document (age, source, etc.). To these ends, methods have been developed for contrast and sharpness enhancement, rectification (including skew and shear correction), super-resolution, and shape reconstruction (for a survey, see [37]) – but there appear to be quite a few open problems.

6 Analysis of Content

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.

However, it should be noted that images found in DLs, since they represent many nations, cultures, and historical periods, tend to pose particularly severe challenges to today's DIA methods, and especially to the architecture of DIA systems, which are not robust in the face of multilingual text and non-Western scripts, obsolete typefaces, old-fashioned page layouts, and low or variable image quality. The sheer variety of document images that are rapidly being brought on line threatens to overwhelm to the capabilities of state-of-the-art DIA systems; this fact, taken alone, suggests that a fruitful direction for DIA R&D is a search for tools that can reliably perform specific, perhaps narrowly defined, tasks across the *full range of naturally occurring documents*. These might include:

1. Does an image contain *any* printed or handwritten text?
2. Does it contain a long passage (*e.g.*, 50 words) of text?
3. Isolate all textual regions, separating them from non-textual regions and background;
4. Identify/segment handwritten from machine-printed text; and
5. Identify script (writing system) and language of regions of text.

This might be called a breadth-first (or *versatility–first*) DIA strategy. Most of these tasks have, of course, already received some attention in the literature. What is new, perhaps, is the emphasis on achieving some level of competency (perhaps not always high) across *orders of magnitude more document image types* than has been attempted thus far.

6.1 Accurate Transcriptions of Text

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. Furthermore, it is rarely possible – even for an OCR expert – 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 [49]). This combination of unreliability, unpredictability, and untrustworthiness forces expensive manual "proofing" (inspection and correction) in all document scan-and-conversion projects that require a uniformly high standard of accuracy. (Of course, if an *average* high accuracy across a large set of documents is needed, existing commercial OCR systems may be satisfactory.)

The open problems here are clearly difficult, urgent, and many, but they are also already thoroughly discussed in the DIA literature (*e.g.*, [41] and [48]).

6.2 Labeling of Structure

DLs would certainly benefit from DIA facilities able to label every part of document structure to the degree of refinement supported by markup languages such as XML. Of course, the general case of this remains a resistant class of DIA problems. However, even partial solutions might be useful in DLs since they would aid in navigation within and among documents, capturing some of the flexibility that keeps paper competitive with DLs. Navigation can be assisted by a wide range of incomplete and even errorful functional labelings for the purposes of, for example, creating indices and overviews (at various levels of detail), jumping from one section to the next, following references to figures, and such.

7 Presentation, Printing, and Reflowing

Paper invite the "spreading out" of many pages over large surfaces. The relative awkwardness of digital displays is felt particularly acutely here. When attempting to read images of scanned pages on electronic displays, it is often difficult to avoid panning and zooming, which quickly becomes irritating and insupportable.

This problem has been carefully and systematically addressed by several generations of eBook design, and progress is being made toward high-resolution, grayscale and color, bright, high contrast, lightweight, and conveniently-sized readers for page images. But even when eBooks approach paper closely enough to support our most comfortable habits of reading, there will still be significant needs for very large displays so that large documents (*e.g.*, maps, music, engineering drawings) and/or several-at-once smaller documents can be taken in at one glance. Perhaps desktop multi-screen "tiled" displays will come first; but eventually it may be necessary to display documents on desk-sized or wall-sized surfaces. The DIA community should help the design of these displays and should investigate versatile document-image tiling algorithms.

In many printed materials, the author's and editor's choice of typeface, type-size, and layout are not merely aesthetic, they are meaningful and critical to understanding. Even if DIA could provide "perfect" transcription of the textual content (as ASCII/Unicode/XML), many critical features of its original appearance may have been discarded. Preserving all of these stylistic details through the DIA pipeline remains a difficult problem. One solution to this problem is, of course, multivalent representations where the original image is always available as one of several views.

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.*, [12]). The intent is to allow imaged documents to be read on a limited-resolution, perhaps even handheld, computing device, without any errors or losses due to OCR and retypesetting, thus mimicking one of the most useful features of encoded documents in DLs. It also holds out the promise of customizable print-on-demand services and special editions, *e.g.*, large-type editions for the visually impaired.

This is a promising start but, to date, document image reflowing systems work automatically only on body text and still have some problems with reading order, hyphenation, etc. Automation of link-creation (to figures, footnotes, references, etc.) and of indices (*e.g.*, tables of contents) would greatly assist navigation on small devices. 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 how this could be accomplished at the present state of the art.

Similar issues arise when users wish to reprint books or articles found in DLs. It should be possible for such a user to request any of a wide range of output formats, *e.g.*, portrait or landscape, multiple "pages per page," pocketbooks, large-type books, etc. In most of these cases, some DIA problem needs to be solved.

8 Indexing, Retrieval, and Summarization

Both indexing and retrieval of document images are critical for the success of DLs. To pick only a single example, the JSTOR DL [29] includes over 12 million imaged pages from over 300 scholarly journals and allows searching on (OCRed) full text as well as on selected metadata (author, title, or abstract field). Most published methods 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). The excellent survey [20] summarized the state of the art (in 1997) of retrieval of entire multi-page articles as follows:

1. at OCR character error rates below 5%, these IR methods suffer little loss of either recall or precision; and
2. at error rates above 20%, both recall and precision degrade significantly.

There is a small but interesting literature on word-spotting "in the image domain." These approaches seem to offer the greatest promise of large improvements in recall and precision (if not in speed). An open problem, not much studied, is the effectiveness of OCR→IR methods on very short passages, such as, in an extreme but practically important case, short fields containing key metadata (title, author, etc.). Many textual analysis tasks (*e.g.*, those that depend on syntactic analysis), whether modeled statistically or symbolically, can be derailed by even low OCR error rates.

8.1 Summarizations and Condensation

There has been, to our knowledge, only a single DIA attack on the problem of summarization of documents by operating on images, not on OCRed text. In this work [15], word-images were isolated and compared by shape (without recognition) and thereby clustered. The cluster occurrences and word sizes were used to distinguish between stop words and non-stop words, which were then used to rank (images of) sentences in the usual way.

This successful extension of standard information retrieval methods into the purely image domain should spur investigation of similar extensions, for example, methods for condensing document images by abstracting them into a set of section headers.

8.2 Non-Textual Content

Non-textual content such as mathematical expressions, chemical diagrams, technical drawings, maps, and other graphics have received sustained attention by DIA researchers, but it may be fair to say that search and retrieval for these contents is at a much less mature stage than for text.

9 Personal and Interactive Digital Libraries

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. All the issues we have mentioned earlier are applicable here, but perhaps there is special urgency in ensuring that all the images are legible, searchable, and browseable. Thus there is a need for deskilled, integrated tools for scanning, quality control and restoration, ensuring completeness, adding metadata, indexing, redisplay, and annotation. An early example of this, using surprisingly simple component DIA technologies informally integrated, is described in [56]. In addition, this might spur more development and wider use of simple-to-use, small-footprint personal scanners and handheld digital cameras to capture document images, with a concomitant need for DIA tools (perhaps built into the scanners and cameras) for image dewarping, restoration, binarization, etc.

In addition, one may wish to detect (near) duplicates, either to prune them or to collect slightly differing versions of a document; the DIA literature offers several effective attacks on this problem (cf. [20]), operating both in the textual and the image domain. Even when document content starts out in encoded form (is "born digital"), document image analysis can still be important. For instance, how might duplicate detection be performed when one of the versions is in PDF format and the other is in DjVu? The common denominator must be the visual representation of the document since, from the point of view of individual (especially non-professional) users, the visual representation will be normative.

Often, users may wish to be able to perform annotation using pen-based input (on paper or with a digital tablet/stylus). A role for document image analysis here could be annotation segmentation/lifting or word-spotting in annotations.

9.1 Interactive and Shared Digital Libraries

As publicly available DLs gather large collections of document images, opportunities will arise for collective improvement of the DL services. For example, one user may volunteer to correct an erroneous OCR transcription; another may be willing to indicate correct reading order or add XML tags to indicate sections. In this way a multitude of users may cooperate to improve the usefulness of the DL collection without reliance on perfect DIA technology. Within such a community of volunteers, assuming it could establish a culture of trust, review, and acceptance, DIA tools could be critically enabling.

An example of such a cooperative volunteer effort, which is closely allied intellectually to the DIA field, is The Open Mind Initiative [58], a framework for supporting the development of "intelligent" software using the Internet. Based on the traditional open source method, it supports domain experts, tool developers, and non-specialist "netizens" who contribute raw data.

Another example, from the mainstream of the DL field, is Project Gutenberg [46], the Internet's oldest producer of free electronic books (eBooks or eTexts). As of November 2002, a total of 6,267 "electronic texts" of books were made available online. All the books are in the public domain. Most of them have been typed in, and corrected (sometimes imperfectly), by volunteers working over the Web. Such databases are potentially useful to the DIA community as sources of high quality ground-truth associated with known editions of books, some of which are available also as images. These collections have great potential to drive DIA R&D relevant to DLs, as well as to benefit from it.

9.2 Offering DIA Tools

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. These could allow, for example, an arbitrary TIFF file (whether in a DL or privately scanned) to be processed, via a simple HTML link, into an improved TIFF (*e.g.*, deskewed). Each such user would be responsible for ensuring that his/her attempted operation succeeded – or, less naively, there could be an independent review. The result would then be uploaded into the DL, annotated to indicate the operation and the user's assurance (and/or the review). In this way, even very large collections of document images could be improved beyond the level possible today through exclusively automatic DIA processing.

References

1. H. Y. Abdelazim and M. A. Hashish. Application of HMM to the recognition of isolated Arabic words. In *Proceedings of 11th National Computer Conference*, pages 761–774, 1989.
2. Association for Information and Image Management, International. 1100 Wayne Avenue, Suite 1100, Silver Spring, Maryland 20910; `www.aiim.org`.
3. R. B. Allen and J. Schalow. Metadata and data structures for the historical newspaper digital library. In *Proceedings of the 8th international conference on Information and knowledge management*, pages 147–153, 1999.
4. A. Amin. Offline Arabic character recognition: The state of the art. *Pattern Recognition*, 31(5):517–530, 1997.
5. A. Antonacopoulos and D. Karatzas. Document image analysis for World War II personal records. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 336–341, 2004.
6. H. Baird. Difficult and urgent open problems in document image analysis for libraries. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 25–32, 2004.
7. H. S. Baird and V. Govindaraju, editors. *Proceedings of the International Workshop on Document Image Analysis for Libraries.* IEEE Computer Society Press, Piscataway, NJ, 2004.
8. V. Bansal. Integrating knowledge sources in devanagari text recognition. *IEEE Transactions on Systems, Man and Cybernetics Part A*, 30(4):500–505, 2000.
9. W. Barrett. Digital mountain: From granite archive to global access. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 104–121, 2004.
10. I. Bazzi, R. Schwartz, and J. Makhoul. An omnifont open-vocabulary OCR system for English and Arabic. *IEEE Pattern Analysis and Machine Intelligence*, 21(6):495–504, June 1999.
11. T. Bray, J. Paoli, C. M. Sperberg-McQueen, and E. Maler. Extensible Markup Language (XML) 1.0 (second edition), 2001.
12. T. M. Breuel, W. C. Janssen, K. Popat, and H. S. Baird. Paper to PDA. In *Proc., IAPR 16th ICPR*, volume 4, pages 476–479, Quebec City, Canada, August 2002.

13. B. B. Chaudhuri and U. Pal. An OCR system to read two Indian language scripts: Bangla and Devanagari. In *Proceedings of the 4th International Conference on Document Analysis and Recognition*, pages 1011–1015, 1997.
14. B. B. Chaudhuri, U. Pal, and M. Mitra. Automatic recognition of printed Oriya script. In *Proceedings of the 6th International Conference on Document Analysis and Recognition*, pages 795–799, 2001.
15. F. R. Chen and D. Bloomberg. Summarization of imaged documents without OCR. *Computer Vision and Image Understanding*, 70(3), June 1998.
16. M. Chew and H. S. Baird. BaffleText: a Human Interactive Proof. In *Proc., 10th IS&T/SPIE Document Recognition & Retrieval Conf.,*, Santa Clara, CA, January 23–24 2003.
17. Unicode Consortium. *The Unicode Standard Version 4.0.* Addison-Wesley, 2003.
18. B. Couasnon, J. Camillerapp, and I. Leplumey. Making handwritten archives documents accessible to public with a generic system of document image analysis. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 270–277, 2004.
19. A. Dillon. Reading from paper versus screens: A critical review of the empirical literature. *Ergonomics*, 35(10):1297–1326, 1992.
20. D. Doermann. The indexing and retrieval of document images: A survey. *Computer Vision and Image Understanding*, 70(3), June 1998.
21. E Ink, 733 Concord Avenue, Cambridge, MA 02138. `www.eink.com`.
22. V. Govindaraju, S. Khedekar, S. Kompalli, F. Farooq, S. Setlur, and V. Prasad. Tools for enabling digital access to multilingual Indic documents. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 122–133, 2004.
23. V. Govindaraju and H. Xue. Fast handwriting recognition for indexing historical documents. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 314–320, 2004.
24. D. Green. The Java Tutorial: Internationalization. `java.sun.com/docs/books/tutorial/i18n/`.
25. Gyricon Media, Inc., 6190 Jackson Road, Ann Arbor, MI 48103. `www.gyriconmedia.com`.
26. G. Harit, S. Chadhury, and H. Ghosh. Managing document images in a digital library: An ontology guided approach. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 64–92, 2004.
27. J. Hochberg, L. Kerns, P. Kelly, and T. Thomas. Automatic script identification from images using cluster-based templates. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 378–381, 1995.
28. L. Hutchison and W. A. Barrett. Fast registration of tabular document images using Fourier analysis. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 253–267, 2004.
29. JSTOR Digital Library. University of Michigan and Princeton University. `www.jstor.org`.
30. D. H. Kim, Y. S. Hwang, S. T. Park, E. Kim, S. Paek, and S. Bang. Handwritten Korean character image database. In *Proceedings of the 2nd International Conference on Document Analysis and Recognition*, pages 470–473, 1993.
31. M. Kim, M. Jang, H. Choi, T. Rhee, and J. H. Kim. Digitalizing scheme of handwritten Hanja historical document. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 321–327, 2004.

32. S. Kompalli, S. Setlur, V. Govindaraju, and R. Vemulapati. Creation of data resources and design of an evaluation test bed for Devanagari script recognition. In *Proceedings of the 13th International Workshop on Research Issues on Data Engineering: Multi-lingual Information Management*, pages 55–61, 2003.
33. M. Kornfield, R. Manmatha, and J. Allan. Text alignment with handwritten documents. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 195–209, 2004.
34. V. Lavrenko, T. Rath, and R. Manmatha. Holistic word recognition for handwritten historical documents. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 278–287, 2004.
35. F. LeBourgeois, E. Trinh, B. Allier, V. Eglin, and H. Emptoz. Document images analysis solutions for digital libraries. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 2–24, 2004.
36. C. H. Lee and T. Kanungo. The architecture of TrueViz: A groundTRUth / metadata editing and VIsualiZing toolkit. *PR*, 36(3):811–825, March 2003.
37. R. P. Loce and E. R. Dougherty. *Enhancement and Restoration of Digital Documents: Statistical Design of Nonlinear Algorithms.* Society of Photo-optical Instrumentation Engineers, January 1997. ISBN 081942109X.
38. H. Ma and D. Doermann. Adaptive Hindi OCR using generalized Hausdorff image comparison. *ACM Transactions on Asian Language Information Processing*, 26(2):198–213, 2003.
39. S. Marinai, E. Marino, F. Cesarani, and G. Soda. A general system for the retrieval of document images from digital libraries. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 150–173, 2004.
40. Microsoft Windows glyph processing. `www.microsoft.com/typography/developers/opentype/default.htm`.
41. G. Nagy. Twenty years of document image analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):38–62, 2000.
42. A. M. Namboodiri and A. K. Jain. On-line script recognition. In *Proceedings of the 16th International Conference on Pattern Recognition*, pages 736–739, 2002.
43. Bureau of Indian Standards. Indian script code for information interchange, 1999.
44. U. Pal and B.B. Chaudhuri. Script line separation from Indian multi-script documents. In *Proceedings of the 5th International Conference on Document Analysis and Recognition*, pages 406–409, 1999.
45. NPES/AIIM PDF – Archive Project. `www.aiim.org/standards.asp?ID=25013`.
46. Project Gutenberg. `promo.net/pg`.
47. S. Rajkumar. Indic typesetting – challenges and opportunities. *TUGboat*, 23(1), 2002.
48. 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.
49. 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.
50. A. J. Sellen and R. H. R. Harper. *The Myth of the Paperless Office.* The MIT Press, Cambridge, MA, 2002.
51. S. Setlur, A. Lawson, V. Govindaraju, and S. Srihari. Large scale address recognition systems – truthing, testing, tools and other evaluation issues. *International Journal of Document Analysis and Recognition*, 4(3):154–169, 2002.

52. S. Shanbhag, D. Rao, and R. K. Joshi. An intelligent multi-layered input scheme for phonetic scripts. In *Proceedings of the 2nd International Symposium on Smart Graphics*, pages 35–38, 2002.
53. S. J. Simske and M. Sturgill. A ground-truthing engine for proofsetting, publishing, re-purposing and quality assurance. In *Proceedings of the 2003 ACM Symposium on Document Engineering*, pages 150 – 152, 2003.
54. B. K. Sin and J. H. Kim. Ligature modeling for online cursive script recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):623–633, 1997.
55. E. H. Barney 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.
56. A. Lawrence Spitz. SPAM: A scientific paper access method. In *Document Analysis Systems II*, pages 242–255. World Scientific, 1998.
57. S. N. Srihari and E. J. Kuebert. Integration of handwritten address interpretation technology into the United States Postal Service remote computer reader system. *Proceedings of the 4th International Conference on Document Analysis and Recognition*, 1997.
58. D. G. Stork. The Open Mind initiative. In *Proc., IEEE Expert Systems and Their Applications*, pages 16–20, May/June 1999. www.openmind.org.
59. K. Summers. Document image improvement for OCR as a classification problem. In *Proc., SPIE/IS&T Electronic Imaging Conf. on Document Recognition & Retrieval X*, pages 73–83, Santa Clara, CA, January 2003. SPIE Vol. 5010.
60. Sun solaris 9 operating system features and benefits – compatibility. www.sun.com/software/solaris/sparc/solaris9_features_compatibility.html.
61. J.-W. Tai, Y.-J. Liu, and L.-Q. Zhang. A model based detecting approach for feature extraction of off-line handwritten Chinese character recognition. In *Proceedings of the 2nd International Conference on Document Analysis and Recognition*, 1993.
62. T. N. Tan. Rotation invariant texture features and their use in automatic script identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(7):751–756, July 1998.
63. The XML version of the TEI guidelines. www.tei-c.org/P4X/CH.html.
64. A.-B. Wang, J. S. Huang, and K.-C. Fan. Optical recognition of handwritten chinese characters by partial matching. In *Proceedings of the 2nd International Conference on Document Analysis and Recognition*, 1993.
65. I. B. Yosef, K. Kedem, I. Dinstein, M. Beit-Arie, , and E. Engel. Classification of Hebrew calligraphic handwriting styles: Preliminary results. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 299–305, 2004.
66. B. Zhang, C. Tomai, S. Srihari, and V. Govindaraju. Construction of handwriting databases using transcript-based mapping. In *Proceedings of the 1st International Workshop on Document Image Analysis for Libraries. (DIAL 2004)*, pages 288–298, 2004.