Pattern Recognition Letters 000 (2015) 1–7

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Disruptive developments in document recognition☆

George Nagy*

Rensselaer Polytechnic Institute, Troy, NY 12180, USA

ARTICLE INFO

Article history:
Received 29 June 2015
Available online xxx

Keywords:
Optical character recognition
Document analysis
Forms processing
Semantic web

ABSTRACT

Progress in optical character recognition, which underlies most applications of document processing, has been driven mainly by technological advances in microprocessors and optical sensor arrays. Software development based on algorithmic innovations appears to be reaching the point of diminishing returns. Research results, dispersed among a dozen venues, tend to lag behind commercial methodology. Some early mainline applications, like reading typescript, patents and law books, have already become obsolete. Check, postal address, and form processing are on their way out. Open source software may open up niche applications that don't generate enough revenue for commercial developers, including poorly-funded transcription of historical documents (especially genealogical records). Smartphone cameras and wearable technologies are engendering new image-based applications, but there is little evidence of widespread adoption. As document contents are integrated into a web-based continuum of data, they are likely losing even the meager individuality of discrete sheets of paper. The persistent need to create, preserve and communicate information is giving rise to entirely new genres of digital documents with a concomitant need for new approaches to document understanding.

1. Introduction

For many centuries, letters, books, maps, drawings and sheet music served to preserve and disseminate ideas and facts. Documents comprising written or printed symbols were the primary means of communicating across time and space. The underlying symbol systems were the crowning achievement of human aspiration for indirect communication. Alphabets, syllabaries and logographs represented spoken language. Specialized notations were developed for music, chess, electrical circuits, architecture, and engineering design. Documents proliferated and spawned the most successful and profitable application of pattern recognition and machine learning. But the migration of information to the web raises troubling questions about the role of documents in the electronic age.

Specialized OCR underlies document image processing applications for engineering drawings, music, maps, and mathematical formulas. A typical E-sized telephone company drawing has about 3000 words and numbers (including revision notices). Musical scores also contain numerals and instructions like *pianissimo*. A map without place names and elevations would have limited use. Formulas and equations abound in digits, letters, alphabetic strings, and mathematical symbols. 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 may provide an opportunity for customization to specialized applications that have not yet attracted heavyweight developers. In the meantime, the conversion of documents containing diverse mixes 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.

Many key components of commercial OCR were foreshadowed decades earlier by academic research. Because all of us know how to read, samples of isolated characters have been frequently used by researchers to test and demonstrate new algorithms. Feature extraction and classification became part of pattern recognition and machine learning, but adaptive and semi-supervised learning has not yet been integrated into commercial OCR. However, most OCR applications require many tasks in addition to character classification. Far more than the sum of its parts, a complete OCR system must demonstrate language and script recognition, colored print processing, column, paragraph and line layout analysis, accurate character/word, numeric, symbol and punctuation recognition, table analysis, adequate language modeling, document-wide consistency, customizability and adaptability, graphics subsystems, effectively embedded interactive error correction, and multiple output formats. Furthermore, specialized systems—for postal address reading, check reading, litigation, and bureaucratic forms processing—also require high throughput and different error-reject trade-offs.

Hand-printed character recognition is either *free-form* or *constrained* by guide lines, character boxes, or "combs" in a drop-out ink

☆ This paper has been recommended for acceptance by S. Sarkar.

* Tel.: +1 518 271 6885; fax: +1 518 276 6261.
E-mail address: nagy@ecse.rpi.edu

http://dx.doi.org/10.1016/j.patrec.2015.11.024
0167-8655/

(invisible to the scanner) that avoids having to separate preprinted boxes from character strokes. Accuracy depends on the training, motivation and consistency of the writer, and on the degree of contextual correction available from the processor's database. Both human and machine legibility depend significantly on the motivation of the writer: a tax return requesting a refund is likely to be more legible than one reporting an underpayment. Immediate feedback, the main advantage of *online recognition*, is a powerful form of motivation. Humans still adapt more readily than machines. Nowadays few people communicate important information via handwriting, except possibly to themselves (as notes or diaries). An exception is countries like China, where much of the population takes pride in calligraphic handwriting.

Among the major remaining applications of handwritten script recognition is the transcription of historical documents created before the spread of typewriters around 1900. So far only experimental systems have surfaced for this purpose, and the necessary scanning and interactive correction make them more expensive than key entry (for historical documents, often by unpaid volunteers).

1.1. Scope of the article

The focus in this article is on disruptive current changes in the objectives and methods of converting ink on paper into a computer-storable, electronically-transmittable and, in a restricted sense, machine-understandable entity. The image-processing aspects of document processing consist of scanning the hardcopy document into a digital image, and converting the image into a symbolic representation that reflects some of its content and appearance (the term d*igitization* may encompass OCR in addition to scanning). Both steps share many aspects with other image processing applications, but the reader must look elsewhere for descriptions and explanations of logo, stamp, seal and signature recognition, text and image compression, cyber security (encryption, watermarking, steganography, CAPTCHAs), scene OCR, license plate readers, and many other related and worthy topics. On-line character recognition, an increasingly important technology in its own right that competes with other text data entry and editing methods in some applications, is also beyond our scope.

The remainder of the Introduction presents a brief history of document image processing and of the industry that it fostered. We also classify the applications reviewed in subsequent sections into those that were so successful that they worked themselves out of a job, those that are still mainstream, and those that are waiting around the corner.

Section 2 describes specialized applications like check and postal address reading that are usually performed by large organizations. Dispersed applications like engineering drawing, map, mathematical notation and music score conversion that are characterized by smaller volumes and fewer resources and are not centrally managed by some organization like a postal service, clearing house, or patent office are discussed in Section 3, as are the niche applications of chess records and calligraphy. Section 4 provides an overview of related downstream information retrieval and document management applications. It explores what can be done with transcribed books, technical journals, magazines and newspapers. It also offers suggestions for finding further information in conference proceedings, technical journals and monographs, and a peek into the future.

1.2. History

In the first half of the 20th Century many inventors were excited by the idea of reading machines for the blind and for automated input to telegraphy. After some false starts, OCR became a competitive commercial enterprise in the 1950's [1]. David Shepard founded Intelligent Machines Corporation. Jacob Rabinow designed the first postal readers. Rapid progress ensued in feature design, classifiers, character and word segmentation, layout analysis, and language modeling. A decade later there were more than 50 OCR manufacturers in the US alone. Their products consisted of scanning equipment and hardwired logic for recognizing mono-spaced OCR fonts and Elite and Pica typewritten scripts one character at a time—eventually at the rate of several *thousand* characters per second. Each of these systems displaced dozens of key-entry operators.

With the advent of microprocessors and inexpensive optical scanners derived from facsimile machines, the price of OCR dropped from tens and hundreds of thousands of dollars to that of a bottle of table wine. Software displaced the racks of electronics. The Optical Character Recognition Users Association (OCRUA) staged popular conferences and published an informative newsletter. By 1985 anybody could program and test their ideas on a personal computer, and then write a paper about it (and perhaps even patent it). Low-end OCR was packaged for give-away with printers.

With enough memory to store an entire page image, word recognition for reading hard-to-segment typeset text became practicable. The value of language models better than letter n-gram frequencies and lexicons without word frequencies gradually became clear. When coding methods (like ASCII for Latin-based scripts) became available for other languages, OCR turned multilingual. This triggered a movement to post all the cultural relics of the past on the Web. Much of the material still awaiting conversion, ancient and modern, stretches the limits of human readability because of paper and ink degradation, and archaic grammar, vocabulary and letterforms. Like humans, OCR must take full advantage of syntax, style, context, and semantics to resolve similar character images like I, l, and 1.

OCR error rates have gradually dropped, but it is useful to remember that a 0.5% error rate still leaves 10 errors on a printed page. Perhaps a more significant contemporary development is detailed tagging of output. Both alphabetic and graphic document components are parsed, and their syntactic characteristics are labeled in the output for the benefit of downstream (non-image) applications (cf. Section 4).

There is, however, little information available to the public about current commercial methods and in-house experimental results. Competitive industries have scarce motivation to publish and their patents may only be part of their legal arsenal. Herbert Schantz's The History of OCR [2] was an exception: it traced the growth of REI (originally Recognition Equipment Inc.), which was one of the major OCR success stories of the 1960's and 1970's. Schantz also told the romantic story 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 many OCR products.

1.3. The industry

Today a few industrial-strength OCR engines dominate the US market: *FineReader* (Abbyy, originally from Moscow), *OmniPage* (product ancestry: Palantir then Calera then CAERE, now Nuance), Readiris (Iris) and *Open Source Tesseract and OCRopus (originally Hewlett Packard's ReadRight*, now sponsored and promoted by Google). Many commercial and open-source software providers incorporate one of the above engines, though some older products are still available. Earlier suppliers, like REI, IBM, Control Data, Burroughs, NCR, Kurzweil, Xerox, ScanSoft, and Recognita, are no longer in this business. However, most of the large Chinese, Japanese, Korean and Arabic OCR providers include English-language subsystems in their products.

Software for handprint and handwriting recognition tends to be too error prone for stand-alone applications. To yield acceptable output, contextual information and an operator interface for correction are embedded in the recognition software. Parascript and A2ia offer

Please cite this article as: G. Nagy, Disruptive developments in document recognition, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.11.024

handprint and hand-writing recognition for postal and administrative documents in several scripts. Remarkably low error rates have been reported on multi-million Chinese and Japanese handwritten character corpora.

Abbyy, Nuance, and Lead Technologies provide software for document format conversion, compression, and for forms-processing, internal mail handling, and other document management applications. Adobe software allows overlaying a searchable (OCR'd) but invisible layer over the rectified page image. Many vendors offer kits for converting non-searchable (image) PDF files into searchable PDF or word-processing formats—essentially an OCR task. Most computer-printer manufacturers also market optical scanners. Although any desktop scanner or multifunction printer can digitize a printed page, some specialized document scanners are described in Section 2.1.

Another segment of the industry consists of service bureaus equipped with high-speed, large-format, microfilm and book scanners, document processing software from multiple vendors, key-entry operators, proofreaders and copy editors. US job shops are now facing stiff competition from lower-wage overseas operators who provide overnight service over the internet.

1.4. Vanishing and emerging applications

Some of the large early applications, like the conversion of US Patents (7 million by 2006) and of the thousands of volumes of edicts by federal, state and local legislatures and courts, have disappeared when searchable versions (mostly produced by key entry) became available. The need for recognizing typewritten material vanished with typewriters, but OCR font and barcode recognition is still used. Shrinking applications include extracting the relevant information from tax returns as all companies and most individuals now file electronically [3].

Technological advances have enabled camera-based OCR. Current smartphone cameras have sufficient resolution for a full page of text. Home applications include capturing financial records (bank checks, invoices), instant language translation, text-to-voice, decision support systems for shopping, and scene OCR (e.g., shop and highway signs).

Historical document processing is tackling older and more degraded documents, some of which are not readable by laymen. Still at the research stage is the automated conversion of old handwritten census, army and legal records, personal diaries and correspondence. The National Archives and Records Administration has compiled extensive military records and maintains the Federal Register. The humungous Congressional Record has been back-digitized to 1983 and posted online. Interest in genealogy fostered the devotion of large resources to the capture and analysis of census records and of birth, marriage and death certificates. The Family Search project of the Church of Jesus Christ of Latter-day Saints deploys teams over the whole world to digitize such material (as well as family and regional history books), engages thousands of volunteers to convert the records to searchable form, and sponsors workshops and research to automate the underlying tasks. Some of the oldest records outside the US are of interest mainly for their calligraphy, ranked by some communities near the top of their cultural heritage.

2. Large, specialized applications

Most of the document processing tasks described here have a direct application to some important ongoing activity where they were previously performed manually—via OCR-font typewriter, keypunch, key-to-disk, or specialized data entry terminal.

2.1. Postal addresses

Among the oldest (since about 1966) and largest applications is postal address reading. The layout analysis problem here is separating sender and destination addresses from background (often advertising) franking, and notices like CONFIDENTIAL. Reading a US destination address is helped by frequently updated directories of all valid mailing addresses and gradually lengthening postal codes for every state, city and street—and even large buildings. Some countries, like Japan, have advanced postal readers in spite of their abstruse address system. Several countries require postal codes in designated boxes.

Incoming and outgoing mail are handled separately: the first is sorted by delivery route, the second, according to the next distribution center. Printed and handwritten addresses are read with an error rate of about 0.5%, and 2% error respectively [4]. The undecipherable images are transmitted electronically to a Remote Encoding Center where the missing or incorrect segment is encoded with very few keystrokes. A barcode with the completed address is added to the envelope to speed up operations at successive sorting stations. US postal volume is down more than 25% from its high of 213 billion pieces in 2006. Furthermore, more and more mail is barcoded at the source to decrease postage and speed delivery, obviating the need for OCR. The number of RECs has decreased from several dozen to one.

2.2. Bank checks

Another large and specialized application is recovering data from checks. The issuing bank's routing number and the account number are read with Magnetic Ink Character Recognition (MICR), which has barely changed since its introduction by Stanford Research Institute in the 1950s. The numeric courtesy amount is compared with the handwritten legal amount and disagreements are corrected manually [5]. Handwritten checks with fancy backgrounds and imprinted amount fields for security are especially hard to read automatically.

High-speed check scanners are used by banks with centralized operations. Since passage of the Check 21 Act in 2003 opened the way for check truncation (image presentment), lower-speed equipment has become available for check capture at teller windows and retailers. Some banks provide free check readers to small businesses for remote entry. Bank checks, at a high of 70 billion in the US in 2001, are being rapidly displaced by electronic payments. Some countries have entirely phased out paper checks.

2.3. Scholarly journals, newspapers, archival records, and cultural heritage

The archival files of many technical journals, magazines, and larger newspapers have already been converted to electronic formats. Like digitized books, most are searchable in spite of uncorrected OCR errors. Some, especially those originally retained only on microfilm, are too degraded for OCR. Conversion of these document image files fall in the rubric of historical document processing. Digitization of the earliest issues of IEEE publications (including thousands of workshop and conference proceedings) remains in progress. An even larger undertaking is the conversion of the holdings of the National Library of Medicine—the largest medical library in the world. Current efforts are aimed at a much fuller representation of technical papers and reports by adding automatically extracted catalog metadata, linking illustrations with the narrative, reverse engineering graphs and tables, and culling, parsing and linking citations and references. Tagging the elements of a document that are significant for a particular application is sometimes called *document understanding*. For example, a properly tagged circuit diagram could be entered directly into a circuit simulator.

The metadata required to classify and find a document is specified by the Dublin Core Metadata Initiative (DCMI) under the aegis of the International Standards Organization (ISO). The Text Encoding Initiative (TEI) Consortium's 1664 page TEI-P5 Guidelines for Electronic Text Encoding and Interchange via the Extensible Markup Language (XML) promotes coherence and consistency in detailed

Please cite this article as: G. Nagy, Disruptive developments in document recognition, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.11.024

tagging of the internal elements of documents [6]. Although TEI also distributes open-source software to assist and validate XML tagging, full automation of the standard is many years away.

Encoding a technical article requires location and identification of titles, subtitles, running heads, page numbers, dates, authors, affiliations, citations, references, figure/table titles and captions, footnotes and footnote references, and many other items. We note, however, that an article or report processed by contemporary OCR software looks as though all the significant components were identified because the layout, fonts and type sizes of the original pages are captured and reproduced or closely approximated in rendering. While this is perfectly adequate for human reading and popular keyword searches, complex queries for data mining entire digital libraries require much more detailed encoding.

Examples of cultural digitization projects include records of the National Archives and Records Administration. NARA already has more than two million digitized copies of its records. Some are processed and posted by Archives partners FamilySearch (open access), Ancestry and Fold3 (via subscription). Other test beds for document interpretation include 5.5 million digitized documents seized during the 1980s Kurdish uprising, 19th Century French military records, and collections of digitized ancient manuscripts in national libraries and museums (some on stone, papyrus, parchment, silk, or palm leaves). The Gallica digital library of the Bibliothèque nationale de France (BnF) offers free download of millions of rare and out-of-print documents.

2.4. Books

Few applications have attracted as much popular attention as Google's 2004 proposal to scan all existing books in a partnership with some of the world's major libraries [7]. By 2013 Google had scanned more than 30 million books. Full text is made available for books no longer subject to copyright, and snippets and some metadata for others. The books are scanned and OCR'd in several dozen languages at the rate of about 1000 pages per hour per machine, but they are not proofread.

Books with simple layouts, modern typefaces and relatively few special symbols can be scanned and OCR'd into searchable formats. However, pages are seldom linked automatically to the table of contents, as they are, manually or via PDF's built-in routines, in all current electronic conference records. Some books are marred by scanning defects like missing, folded or upside-down pages and by OCR errors [8]. Many of the available high quality eBooks, including over 47,000 books of the Gutenberg Project, were keyed in manually.

2.5. Forms and tables

Forms are used for collecting information. They are also called *bureaucratic forms*, *office forms*, *official forms* or, more specifically, *tax returns*, *invoices, insurance applications*, *betting tickets*. Forms were sometimes published in newspapers with a request for reader feedback. The first commercial form reader, for reader subscriptions, was installed at *Readers Digest* in 1955. By 1959 it had read its billionth character. The IBM 1975 page reader—for well formatted and spaced lines of typewritten and printed identification numbers, first and last names, and contributed dollar amounts—was delivered to the Social Security Administration in 1966 with a price tag of over three million dollars.

Many organizations still maintain websites with downloadable forms, but online-fillable web forms have quickly replaced typed and hand-printed forms. Security measures may, however, prevent filled-out forms from being downloaded or saved by the client. Although over 75% of the four million Medicare claims per day arrive electronically, medical insurance remains among the largest forms processing applications.

The principal elements of a form are *labeled fields* demarcated by line art or color. The blank spaces for entering information may include horizontal lines, combs, or other character separators. Fields may also be grouped by line art or color at one or more levels to facilitate entering the required information. Most form recognition software offers templates for customization to specific layouts and preprinted keywords and for verification against information stored in the form processor's database. Forms often solicit redundant information for error detection or correction.

The labels are preprinted in or near the blank where the information is to be entered. Labels may range from a single word to an entire paragraph, possibly in several languages. Forms may also contain check boxes. Line art and labels may be printed in a *drop-out color* invisible to the designated scanner. *Mark sense forms* represent an extreme combination of drop-out ink and check boxes.

In addition to a *Form Name* like "Application for Driver's License", professionally designed forms usually have a, *Form Number*, *Version Number*, or *Date of Issue* and sometimes a unique barcoded identifier. Forms may also show instructions, organizational affiliation (including logos), source, signature lines, spaces for stamps, and advertising. The preprinted instructions may include lists or tables: e.g., state sales tax rates.

Form configurations range from the very simple, like forms for recording tournament chess games, to the outright recondite (like some IRS tax forms). Even simple forms may have dozens of repetitive fields and continuation pages. Among the overwhelming advantage of web forms is their ability to stretch to accommodate varying amounts of information and their ease of correction as opposed to correcting errors on multiple carbon or self-inking copies.

A special case of forms processing is the conversion of hospital patient jackets in the drive toward universal electronic medical record-keeping. Unlike mass conversion of financial and hospital records, historical forms projects are perennially short of resources. Old census forms and birth/marriage/death records are often transcribed manually by volunteers. Research on historical form conversion is, however, gaining popularity as other applications disappear.

In contrast to forms, tables are used for the dissemination rather than the collection, of information, and they seldom appear as standalone documents. Printed tables in books, technical design manuals, journals, patents, and even newspapers (stock market, election or baseball statistics) contain an enormous amount of quantitative information [9]. When such material is transplanted to the web, the tables are often included only in raster image form, or in a simple tabular format like that of a word processor. However, what is wanted for massive data analysis is the more structured representation offered by relational databases that can be queried with SQL, or the newer "ontological" Resource Description Framework (RDF) triples queried with SPARQL. This requires software that discovers the relationship between the (possibly multi-line, multi-category) row and column headers that index every value cell of the table. Research in this direction is conducted by Google, Microsoft, HP and several academic groups.

3. Dispersed applications

This section describes document processing tasks with less economic motivation and urgency than the above. Operations are generally dispersed among many heterogeneous organizations.

3.1. Maps, engineering drawings and schematic diagrams

Mobile-accessible web apps are displacing atlases, topographic maps, road maps, marine charts and globes. Modern maps, like almost all contemporary documents, are produced by computer, often as byproducts of data for Geographic Information Systems (GIS) [10]. The common map file formats are either easily rendered color raster

Please cite this article as: G. Nagy, Disruptive developments in document recognition, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.11.024

files (e.g. DRG, ADRG and RPF) or more complex vector formats with layers of point, polyline and polygon entities (USGS DLG, OpenGIS GML, and Esri Shapefile). There is some interest in rapid automated processing of local hardcopy maps for expeditionary forces in remote military theaters.

Historical map conversion, still primarily at the research stage, does not necessarily aim for application-oriented formats. Specific projects focus on locating buildings and boundary lines in cadastral maps, roads on highway maps, soil types in soil maps, minerals in geological maps, contour lines and waterways in old topographic maps, street names and street lines in city maps, and soundings in hydrological charts.

Digitization and interpretation of engineering and architectural drawings, circuit schematics, wiring diagrams, and other graphical manifestations of technical design are at a stage similar to that of maps because they too have arcane conventions for layout and symbology. Also, like maps, they have a long life time. All current production is by computer-aided drafting and design (CADD). However, some of the DC-3 aircraft launched in 1935 are still flying, and so are the 1970's Boeing 747 Jumbo Jets (reputedly the last commercial airplanes manufactured according to hand-drafted plans). Electric and gas lines, railroad tracks and roadways, and many buildings have even longer operational lifetimes. Most of these artifacts require reviving their outlived documentation in electronic form for updates and maintenance.

Ironically, some of the early CAD software is no longer available or functional, so engineering drawings produced by these systems must be scanned for conversion to modern file formats. The conversion of 3-D machine drawings is particularly demanding and has been subject to decades of research. Raster-digitized drawings produced by obsolete software are often off-shored for re-entry via contemporary stylus-based drafting systems.

3.2. Equations, formulas and chemical diagrams

Ordinary text consists of a linear sequence of symbols, but in mathematics the relative size and position of the symbols is meaningful. Equation recognition has long been an established field of research with its own workshops, conference sessions and competitions [11]. The advent of stylus computers was largely responsible for the recent flourish of online equation recognition, which allows entering a formula in the way learned in high school, without resorting to a math editor like MathType or AMS Math (the American Mathematical Society's LaTeX equation editor). The 2014 Competition on Handwritten Mathematical Expression Recognition (CROHME) compared performance on isolated math symbol recognition, math expression recognition, and matrix recognition.

As in the case of text, tables and maps, labeling the symbols and keeping track of their location is sufficient for printing or displaying them for a human reader. Such a representation is not, however, expressive enough for numerical analysis, symbolic algebra or simulation programs like Maple and Mathematica. The ultimate goal of research in equation recognition must therefore include the conversion of printed and hand printed formulas to mathematical mark-up languages like Open Math and MathML.

Manipulation of structural formulas in chemistry and predicate calculus expressions in formal logic faces similar problems. Drug research firms have long ago converted to computerized operations, and Artificial Intelligence programs have manipulated logic formulas for many decades. Nevertheless, there have been only scattered efforts to convert images of written and printed notations of organic chemistry and formal logic to machine-manipulatable form. The searchable Google Book version of Principia Mathematica illustrates the difficulty of such material.

3.3. Music

Music has a prominent place in contemporary culture, entertainment, and commerce. The conversion of scanned musical scores has been a research objective since the early days of music synthesizers, and music OCR programs have been available 2 decades. Like the language of mathematics, that of music is essentially two-dimensional, and its transcription requires graphics routines in addition to conventional OCR. Current programs can convert scores to playable and editable Musical Instrument Digital Interface (MIDI) notation. Participants at a recent conference could listen to songs generated from sheet music captured in situ by the presenter's smartphone camera [12].

3.4. Niche applications

Researchers have demonstrated the recognition of many symbol systems. Some of these, like the above music app, will surely soon reach commercial maturity. Reading cook books may lead to automating grocery orders and perhaps even chefs. Interest in genealogy led to experiments on reading obituaries, birth and marriage notices, Jiapu (Chinese family histories), and photographs of tombstones.

A particularly interesting demonstration was "Reading Chess", by Baird and Thompson [13]. The program devised by the authors was able to reconstruct complete chess games from several volumes of the poorly printed *Chess Informant* by taking advantage of the codified rules of chess and of the constrains imposed by each successive move. It even found some errors in tournament records.

In China there has been recent interest in producing advertisements in ancient calligraphic fonts. The study of calligraphy is also a popular pastime. Researchers have reported attempts to recognize the calligraphic *style* (e.g. seal, clerical, standard, cursive, and running) of ancient manuscripts. The 32-bit GB 18030 and Unicode 7.0 contain codes for over 50,000 Hanzi logograms, all but ~6000 of which seldom if ever appear in modern documents. Many Chinese characters have been transplanted to early Japanese, Korean and Vietnamese writing systems.

Although low-end document image processing applications are often packaged with other computer software and peripherals, they have not seen wide home use. That is likely to change as ever more powerful smartphones with high resolution cameras saturate the market. Personal document processing differs from mass conversion in amount of data, customization of desired output, range of tasks, individual skill and usage levels, and the jarring effect of unpredictable system responses. It requires (1) *a black-box system*, like a spreadsheet, that is tunable by lay users, (2) *predictability*, which allows the user adapt to the system because the machine communicates and illustrates the source of errors, and (3) *interaction based on meaning rather than appearance*, similar to that between humans discussing an obscure document.

4. End notes

Pervasive use of many important document applications depends on the economic trade-off between the development costs of automation and computer-assisted key entry. While the most valuable documents have been and are still being transcribed by hand, the transcription of continually refreshed (but dwindling) streams of hardcopy documents like postal envelopes, bank checks and medical insurance forms has been successfully automated. Clinical patient records containing a mixture of formatted and unformatted print, hand-print, handwriting, graphics, imagery and audio are still largely converted manually in response to mandates for digitizing medical records. Socially and commercially less urgent document transcription applications also await further progress in automation.

Please cite this article as: G. Nagy, Disruptive developments in document recognition, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.11.024

The semantic and cognitive dimensions of document perusal need more work (e.g. reading order and beyond). These aspects become more and more important since the users do not necessarily distinguish between the textual and image representation of the content. Furthermore, handwritten text recognition in English and print recognition in some non-Latin scripts remain inadequate. Lack of the critical commercial segment may account for the lack of quality OCR in many languages.

Some significant downstream applications that require digitized documents are listed in Section 4.1. Sources of further information about the material presented in this article are listed in Section 4.2. In Section 4.3 we speculate about the on-going shift in the very nature of documents. Section 4.4 is an attempt at a "take-away".

4.1. Downstream technologies

Although Information Retrieval (IR) is not generally considered part of Document Image Analysis (DIA) or vice-versa, the overlap between them includes "logical" document segmentation, extraction of tables of contents, linking figures and illustrations to textual references, and word spotting. A recurring topic is assessing the effect of OCR errors on downstream applications. Common IR methodologies that follow DIA or key entry are keyword search, relevance feedback, document categorization, summarization, authentication, concordance, literary or forensic author identification, and duplicate, forgery and plagiarism detection.

In addition to information retrieval, downstream applications include mailroom automation, advertising and political campaigns targeted to individuals' reading and writing, decision support systems (medical, financial, legal, utility, military, and political), and text re-use (for user manuals or advertising). *NewsStand* makes use of web crawlers, gazetteers, spatial synonyms, toponym resolution, and geotagging to quasi-instantaneously gather all the news pertinent to a location pointed-to on a zoomable world map [14].

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 the correction of the contrast and geometric distortions of camera-captured document images goes well beyond what is required for scanned documents.

4.2. Additional sources of information

Information relevant to the topics discussed is readily available online (though some only via academic, corporate or library subscriptions) in conference proceedings, technical journals, specialized collections, and monographs. The biennial International Conference on Pattern Recognition (ICPR) has a document processing stream, and the smaller International Conference on Document Analysis and Recognition (ICDAR), held in alternate years, is entirely dedicated to this subject. More detailed experimental reports can be found in the proceedings of the even smaller Document Analysis Systems (DAS) workshops and in those of the annual IST/SPIE Document Recognition and Retrieval (DR&R) conferences. More specialized workshops, on historical image processing, equation recognition, graphics recognition, handwriting analysis, and camera-based OCR, are organized around ICPR and ICDAR. The International Conference on the Frontiers of Handwriting (ICFHR) and Graphonomics (IGS) meetings concentrate on the eponymous topics, the ACM Document Engineering Conference (DocEng) on document semantics, and the annual Family Search workshops in Salt Lake City on genealogical applications.

The *International Journal of Document Analysis and Recognition* (IJDAR) is devoted to document research. However, relevant articles occasionally appear in the *IEEE Transactions on Pattern Recognition and Machine Intelligence* (PAMI), the *journal Pattern Recognition* (PR), *Pattern Recognition Letters* (PRL), and the *Journal of Electronic Imaging* (JEI).

Among edited collections the largest are the two-volume *Handbook of Document Image Processing and Recognition* [15] (also in electronic form), and the still useful *Handbook of Character Recognition and Document Image Analysis* [16]. Many of the conference programs are revised for publication by commercial publishing houses. *Machine Learning in Document Analysis* and Recognition [17], and *Advances in Digital Document Processing and Retrieval* [18] contain individually solicited expert contributions. For information about publishing practices and vocabulary, *The US Government Printing Office Style Manual* or *The Chicago Manual of Style* (both available online) may be consulted.

A 2011 monograph by Ferilli [19] provides an excellent introduction to the subject and some chapters on its more arcane aspects. *Character Recognition Systems* [20] offers extensive introductory material on feature extraction and pattern classification, and several case studies. Still relevant are parts of the entertaining and informative *Managing Gigabytes* [21] and of two 1999 books on character recognition, by Rice et al. [22] and by Mori et al. [23], respectively.

4.3. Future prospects

In spite of ubiquitous phone cameras and tireless satellite imagers, documents remain the most common digital images. Journalists and advertisers fill newspapers and magazines, historians indite chronicles, economists explain the stock market, geographers and Google operatives map the world in ever greater detail, diarists record their observations and opinions, bloggers air them, and novelists compete for virtual shelf space on Amazon.

We pass through life leaving a broad trail of documents. We first acquire a birth certificate, then write essays and take tests at school, fill out census, income tax, license and insurance application forms, send birthday cards and checks, sign credit card invoices, accumulate fat folders at our doctor, clinic and hospital, and eventually exit with estate tax forms, an obituary and a death certificate. Document processing offers the possibility, for better or worse, of analyzing and preserving for future generations all of the above.

More and more documents are, however, both generated and read only by computers. They will all be tagged eventually, in whole and in part, in conformance with the dictates of the semantic web [24]. A glimpse of the power of machines able to read, remember and cognitively manipulate an enormous number of documents is the Jeopardy world champion Watson supercomputer, whose powers are now being turned to medical and business applications [25,26]. The knowledge now trapped in individual documents will inevitably migrate into a ubiquitously accessible continuum of facts and opinions. Evidence of this transition is already emerging in the form of micropublications that provide a succinct and information-technology-friendly format for scientific communication [27].

Wearable imaging devices offer the intriguing prospect of capturing and retaining in a personal database every piece of text that one has ever read or even glance at, in books, newspapers, street signs, advertisements, smartphones and computer displays. Systems that enable one to recall the original context of all the fragments of text that one has seen may be a suitable target for the next phase of DIA research.

4.4. Summary

Some of the early drivers for the development of OCR to eliminate the cost of key-entry of patents and judiciary records have disappeared, but almost all document processing applications related to image processing still rely on character recognition.

Please cite this article as: G. Nagy, Disruptive developments in document recognition, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.11.024

Progress in computing and scanning technologies and in algorithm design has expanded the scope of document processing from simple typescript to complex magazine pages and graphic documents. Word error rates on clean, high-contrast print were reduced to the levels expected from professional proofreaders. Commercial and government adoption targeted mainly high-volume transactional applications like checks, postal envelopes and bureaucratic forms. OCR systems, albeit of variable efficacy, became available for almost every language and script and we are well on our way to converting all existing books to electronic files.

Except for genealogy-related records, historical documents have attracted less funding. Nevertheless significant research in document image analysis is devoted to improving the accuracy, speed and scope of the conversion of complex, low-contrast, degraded historical documents in a way that preserves the essential connection between the textual and visual aspects of these artifacts.

While image processing techniques are already widely applied to the conversion of maps, schematics and engineering drawings, these applications still require costly operator interaction for acceptable accuracy. The digitization of medical records, a huge but transient problem, illustrates the need for more integrated approaches to document recognition.

At the opposite end of the spectrum from mass document processing applications, the new generation of digital cameras, including wearable devices, opens the way for mobile personal text-image capture, transcription, and lifelong retrieval.

The increase in the volume of digital documents due to fully or partially automated production and conversion resulted in a corresponding increase in the scope and power of downstream information retrieval and document management applications. Perhaps some types of documents will lose their individuality and unitary nature through integration into quasi-universal information depositories organized along principles entirely different from those of traditional archives and libraries. At the same time, documents themselves are evolving in response to their digital environment. E-books, dynamic web tables, interactive forms, manipulatable illustrations, blogs, wikis, micropublications, and ontologies are entirely new genres. What an exciting time for research on the tools for human communications!

Acknowledgments

Valuable advice from Desirée Butterfield, Archivist/Librarian, Fogler Library, University of Maine, is gratefully acknowledged. The author also appreciates the compliments and insightful suggestions of three knowledgeable referees.

References

[1] M.E. Stevens, Automatic character recognition, a state of the art report, National Bureau of Standards Technical Note 112 (May 1961), https://archive.org/details/automaticcharact112stev.
[2] H.F. Schantz, The History of OCR: Optical Character Recognition, Recognition Technologies Users Association, 1982 0943072018, 9780943072012.
[3] Internal Revenue Service, "U.S. Taxpayers efiled more than 82 million returns so far in 2015, http://www.efile.com/efile-tax-return-direct-deposit-statistics/.
[4] PostalReporter NewsBlog, USPS To Close wichita remote encoding center, posted Feb 20 (2013) http://www.postal-reporter.com/blog/usps-to-close-wichita-remote-encoding-center-797-employees-affective/, (accessed 18.10.15).
[5] Parascript, CheckUltra, http://www.parascript.com/checkultra/
[6] Text Encoding Initiative, P5 Guidelines, http://www.tei-c.org/Guidelines/P5/, Nov. 11, 2014.
[7] D-S Lee, R. Smith, Proceedings of 2012 10th IAPR International Workshop on Document Analysis Systems, 2012, pp. 115–119.
[8] K. Goldsmth, The Artful Accidents of Google Books, The New Yorker, 2013 Dec. 4.
[9] A. Halevy, P. Norvig, and F. Pereira, The unreasonable effectiveness of data, IEEE INTElliGENT SYSTEMS. March/April 2009.
[10] M.F. Goodchild, Twenty years of progress: GIScience, J. Spatial Inf. Sci. 27 (2010) 3–20 Number 1 Jul.
[11] D. Blostein, R. Zanibbi, Springer-Verlag, 2014, pp. 679–702.
[12] H-N Bui, I-S Na, S-H Kim, Staff line removal using line adjacency graph and staff line skeleton for camera-based printed music scores, in: Proc. 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 2014, pp. 2787–2789.
[13] H.S. Baird, K. Thompson, Reading chess, IEEE Trans. Pattern Anal. Mach. Intell. 12 (6) (1990) 552–558.
[14] H. Samet, et al., Reading news with maps by exploiting spatial synonyms, C. ACM 57 (10) (2014).
[15] D. Doermann, K. Tombre (Eds.), Handbook of Document Image Processing and Recognition, Springer Reference, 2014.
[16] H. Bunke, P.S.P. Wang (Eds.), Handbook of Character Recognition and Document Image Analysis, World Scientific, 1997.
[17] S. Marinai, H. Fujisawa (Eds.), Machine Learning in Document Analysis and Recognition, Springer, 2007.
[18] B.B. Chaudhuri, S.K. Parui, Advances in Digital Document Processing and Retrieval, World Scientific, 2014.
[19] S. Ferilli, Automatic Digital Document Processing and Management, Springer, 2011.
[20] M. Cheriet, N. Kharma, C-L Liu, C.Y. Suen, Character Recognition Systems, Wiley-Interscience, 2007.
[21] I.H. Witten, A. Moffat, T.C. Bell, Managing Gigabytes, Van Nostrand Reinhold, ISBN-13:978-1558605701ISBN-10:1558605703, 1994.
[22] S. Rice, G. Nagy, T. Nartker, Optical Character Recognition, Kluwer 1999.
[23] S. Mori, H. Nishida, H. Yamada, Optical Character Recognition, Wiley – ISBN: 978-0-471-30819-5, 1999.
[24] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, scientific American, May 2001.
[25] R. Wanjiku, IBM Pushes Watson's Role in Healthcare, IT World, 2014 , http://www.itworld.com/article/2698674/hardware/ibm-pushes-watson-s-role-in-healthcare.html.
[26] M. Gaynor, G. Wyner, A. Gupta, Dr. Watson? Balancing automation and human expertise in healthcare, in leveraging applications of formal methods, verification and validation. Specialized Techniques and Applications, LNCS 8803, 2014, pp 561-569.
[27] T. Clark, P.N. Ciccarese, C.A. Goble, Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications, J. Biomed. Semant. 5 (2014) 28.

Please cite this article as: G. Nagy, Disruptive developments in document recognition, Pattern Recognition Letters (2015), http://dx.doi.org/10.1016/j.patrec.2015.11.024