Optical Character Recognition

Document Image Analysis: What Is Missing?

George Nagy

ECSE, RPI, Troy, NY, USA 12180-3590, nagy@ecse.rpi.edu

Abstract

The conversion of documents into electronic form has proved more difficult than anticipated. Document image analysis still accounts for only a small fraction of the rapidly-expanding document imaging market. Nevertheless, the optimism manifested over the last thirty years has not dissipated. Driven partly by document distribution on CD-ROM and via the World Wide Web, there is more interest in the preservation of layout and format attributes to increase legibility (sometimes called "page reconstruction") rather than just text/non-text separation. The realization that accurate document image analysis requires fairly specific pre-stored information has resulted in the investigation of new data structures for knowledge bases and for the representation of the results of partial analysis. At the same time, the requirements of downstream software, such as word processing, information retrieval and computer-aided design applications, favor turning the results of the analysis and recognition into some standard format like SGML or DXF. There is increased emphasis on large-scale, automated comparative evaluation, using laboriously compiled test databases. The cost of generating these databases has stimulated new research on synthetic noise models. According to recent publications, the accurate conversion of business letters, technical reports, large typeset repositories like patents, postal addresses, specialized line drawings, and office forms containing a mix of handprinted, handwritten and printed material, is finally on the verge of success.

1 Introduction

As a thirty-five year veteran, I feel entitled to air some insights or prejudices nurtured in the course of years of patient labor in the vineyards of Optical Character Recognition and Digital Image Analysis. I make unsupported statements and unfounded claims without apology. I focus heedlessly on picayune issues like representative samples at the expense of important contributions like new features. I vent my indignation over fuzzy propaganda and wax irascible about persistent failures. To avoid misattribution of inflammatory ideas to innocent parties, I dispense altogether with citations and references, even to my own rambling publications.

I propose to explore continuity and change in research paradigms and real-world applications in OCR and DIA. To this end, I try to differentiate innovative solutions that make me exclaim "I wish I had thought of that!" from the application of new technology to old problems. Of course, in many instances, someone did think of *that* long ago, but lacked the means to test the idea, or the time to write it up. As in other fields, many good ideas are forgotten and reinvented again and again.

Commercial OCR and DIA (the two have always been inseparable) have just celebrated their 40th birthday and may therefore be rightfully suspected of nearing a mid-life crisis. Complaints abound. Researchers in other disciplines, even those in other areas of pattern recognition, consider OCR a long-solved problem. Many regard document image analysis as *infra dig* compared with medical, astronomical, and satellite image analysis and computer vision. Scholars, bureaucrats, lawyers, engineers or geographers seldom make use of OCR and DIA. While document imaging already represents a huge market, only a tiny fraction of the current applications requires techniques beyond the "trivial" processes of acquiring, indexing, compressing, storing, transmitting, retrieving, and displaying thousands of full-page images. Potentially, these files are all grist to our mill.

I first review the earliest applications of pattern recognition to printed matter, and trace the gradual shift to more and more demanding tasks. Then I list some old ideas that have stood the test of time before mentioning techniques that have been adopted or considered seriously only recently. Because the evaluation of OCR and DIA systems has recently become such an active area, I dedicate a short section to this topic. I conclude with the mandatory expressions of faith, optimism, and belief in progress.

2 Old Applications

The easy problems were solved in the sixties and early seventies, but the economics and methods of implementation have changed. Systems that used to require expensive electromechanical or CRT scanners and hardwired special purpose processors now run on PCs using scanners and paper transports that are first cousins to inexpensive fax machines. Nevertheless, the most successful automated data-entry applications still involve specially designed fonts and rigidly formatted documents, where the document preparation is controlled by the organization that must eventually read it (*turn-around forms*). OCR-A and OCR-B fonts, which have both undergone several rounds of revision and expansion by standards committees, are still alive. The phrase *document reader,* which used to refer to reading forms, now means a machine that reads entire pages of text.

The multi-million dollar IBM 1975 that was installed at the Social Security Administration headquarters in Baltimore in 1965 read full-page, "omni-font" reports of social security numbers, names, and earnings. It chugged along for more than twenty years, reading millions of typed and line-printer generated pages. The IBM 1975 had thresholds for rejecting both individual lines and whole pages, and it rejected plenty. The resulting low error rate was improved further downstream by automated cross-checking against the SSA's master-file of one million names.

From the point of view of pattern recognition, the essential aspect of form reading is that most fields contain at most a dozen characters of interest (e.g. name, address, account number, amount). Thus the system can run at a relatively high character reject rate and still recognize all the characters on most fields. At a high reject rate, expensive substitution errors are very rare. Any field on which even a single character is rejected is submitted to manual handling. In the last decade, the availability of sufficient digital storage to retain and display the entire field has increased the efficiency of operator intervention enormously. The form that contains the error no longer needs to be physically accessed: scanning, recognition, and post-processing can be separated in time and space. Current form readers can process printed, imprinted (stamped), and handprinted characters on the same form, as is typical on credit card slips.

Another important application introduced in the sixties was postal address reading. The first machines read only *outgoing mail* (city, country and postal code). In Europe and the United States, most of the addresses were typed or printed, but in Japan the postal codes were handprinted in preprinted boxes. Printed and handwritten mail were separated before the machines took a crack at it. Some machines were later modified to read *incoming mail* (street addresses). The current machines find and read the entire address block and spray a bar-code on the envelope (including, in the US, two extra digits, for a total of eleven, for delivery sequencing). Accuracy is enhanced by reference to a postal address database. The US database takes about 300 megabytes and is updated weekly. As in the case of form readers, postal readers pay their way even if they reject 30% of the mail. Postal research, spearheaded in the US by CEDAR, has also produced many innovative byproducts for other applications. However, reading 95% of the first-class mail is still five years away, just as it was in the sixties.

A bane of early postal readers, blotchy Addressograph plates,* has virtually disappeared. Another application that was displaced in the seventies (by key-to-disk) was retyping documents, on a typewriter with an OCR font, as a substitute for keypunching. More importantly, the replacement of office typewriters and chain printers by 300 dpi laser printers, and of thermal copiers by intelligent reprographics, did wonders for OCR error rates on office documents. (But typewriters and dot-matrix printers are still important in some form-processing applications.) We must note also the remarkable difference between the image quality of CCD scanners and that of the flying-spot scanners of yore.

3 New Applications

Among the new DIA applications, the most compelling is the World Wide Web. As anyone who has surfed the Internet can testify, it is an ocean of meta-information and pointers to documents still on paper. Some current authors do post their latest drafts, converted with minimal human intervention from TeX to HTML (the Web *lingua franca*), but the bulk of the technical material is reference information: bibliographies, directories, and catalogs (some already obsolete). Housekeeping is rudimentary. The cost of keying-in archival technical information, such as the contents of the last ten years' accumulation of the 2000-3000 largest-circulation periodicals and conference proceedings, is staggering.

Nevertheless, the WWW is likely to host the first large-scale digital libraries. A recent call for consortium proposals by the National Science Foundation, the Advanced Research Projects Agency, and the National Aeronautical and Space Administration drew so many applications that NSF was hard-pressed to find knowledgeable referees who were not affiliated with any applicant. Five projects, led by major research universities, were funded at about five million dollars each. However, only a small part of this funding addresses conversion from paper to electronic form.

Other important DIA targets are the repositories of the world's bureaucracies. Government agencies yearn to convert their bulging files to computer readable form, ready to be searched, indexed, annotated, and shuffled from civil servant to civil servant. We cite three such applications in the United States that have triggered significant DIA activity. The first is the Department of Energy's Licensing Support System, which contains all the documents relevant to the disposition of spent nuclear fuel. Some of these papers may be subpoenaed in lawsuits in the year 3000! The legal requirements to retain access to the information - recently estimated at forty million pages - spawned the Information Science Research Institute at the University of Nevada.

Increasing demands on the Bureau of Census for timely demographic information (primarily for marketing purposes) triggered the first and second competitions on HCR (handprinted character recognition). The tests were conducted by the National Institute of Standards and Technology whose competence in OCR can be traced at least to Jacob Rabinow's 1952 "Rapid Selector" for upper-case typewritten characters. The training material for the first contest consisted of forms written by Census employees. The contestants were tested on similar forms filled out by students. The second test was based on three fields related to employment from 1980 census forms (paper and microfilm). Many of the entries were from Europe, but the volume of data and the rigid reporting requirements precluded "amateur" participation. The analysis of the results of these massive tests brought to light many issues related to sampling, error metrics, recognition confidence markers, reject/error trade-offs, and the use of subject-specific context in the form of lexicons.

The third example is the current Internal Revenue Service's initiative to automate processing of individual income tax forms. In a pilot project, many forms are already scanned and stored in compressed bitmap form, which reduces to a few weeks the delay in making them available to local IRS bureaus for detailed audits. The OCR effort is concurrent with the continued encouragement of electronic filing, so fewer forms will be processed than are now manually entered. The system development, which is scheduled to last about five years, is to be shared by large military and aero-space contractors. The need for accurate form classification, form parsing, printed and handprinted character recognition, internal and external contextual constraints, and tight quality control constitute a challenge that eclipses in scale the mammoth SSA system of the sixties.

Most of the large legal reference libraries have already been key-entered for on-line access (*Lexis, WestLaw*), but systems including OCR components are marketed for trial-document management. Besides large, concentrated projects, there are a number of successful mom-and-pop operations. Examples include systems to scan and file visiting cards, read classified advertisements to obtain accurate price ranges for used cars, and convert music notation for publishing new arrangements.

Inexpensive desk-top scanners and OCR programs are reviewed regularly and enthusiastically in personal computing magazines. These systems interface directly with the most popular word processors, so errors can be readily corrected. On well-spaced, high-contrast natural-language text in a normal point size, the error rate is no higher than that of an average typist. It increases significantly on low-contrast, crowded copy, on tables and equations, and non-text alphanumeric strings (because most systems make use of letter n-gram tables or lexicons). Recently several OCR vendors have added fax-reading capability: as expected, the error rate at the normal fax resolution is much higher than at the usual 300 dpi. Many OCR systems offer versions for several languages. According to the current wisdom, individual OCR users don't want to train their machines for new typefaces, but provisions must be included for incorporating special symbols. The number of OCR systems sold annually is still insignificant compared to the number of PCs, and in the United States most of the leading OCR vendors are small, specialized companies.

Among the few examples of line-drawing systems in actual use, a shining example is the cadastral system developed by the late IBM Scientific Center in Rome. In other applications, tracing drawings and maps on a digitizing tablet is gradually being replaced by tracing the lines, using a mouse or "electronic ink," on a raster-digitized version of the drawing. Some systems offer a vectorizing capability: any errors can be easily corrected by the operator using computer-aided drafting commands. To the best of our knowledge, none of these systems have a useful OCR component, and labels, captions and dimensions must be keyed in. However, the most time-consuming aspect of label entry, tying the label to an appropriate graphic entity or coordinate position, has been greatly accelerated by point-and-click operation or automatic high-lighting of consecutive fields.

4 Old Ideas

In this section I list some of my pet aversions as a referee. Many of the techniques mentioned have a lively and productive past and perhaps an honorable and secure future, but they tend to make me nod off within a few paragraphs. Because most are quite central to DIA and OCR, they are revisited again and again. The graduate students who implement them feel compelled to write up each new wrinkle, and thesis advisors (like me) complacently affix their names. The ultimate blame lies with promotion and tenure committees who would rather count than read.

Among preprocessing methods, I single out global thresholding. Almost any binarization method works for high-contrast documents, and no *global* thresholding can be effective on faded or colored copy. A good scanner is worth three times its weight in preprocessing algorithms.

Text-nontext filters based on the spatial spectrum and related orthogonal expansions, morphological operations, and fractal dimensions are solutions in search of a problem. Commercial zoning algorithms do a surprisingly good job of locating text. Methods that do not take into account typesetting rules and layout conventions gerrymander each page into a patchwork guaranteed to bedevil any downstream program.

I am tired of ad hoc format analysis with built-in rules for narrow, specific types of documents. After the first few demonstrations that one can write a program to separate the titles of articles from photo captions, or subheadings from citations, the novelty fades. Writing the rules in a new "high-level" syntax instead of Lisp or Prolog or C does not warrant sacrificing yet another graduate student.

Preprocessing followed by character segmentation followed by feature extraction followed by classification followed by context correction makes for a nice flow chart, but there must be more interesting ways of arranging these blocks. Isn't it time to introduce a few feedback loops?

While speaking of features, please don't send me any more papers about hand-crafted "topological" features or "new" orthogonal expansions, no matter how many hundreds of characters you offer that demonstrate their palpable superiority. For a specific set of shapes, any hacker can think of dozens of workable, easily programmed features. They won't be any better than the hundreds that have been tried (although they may not be much worse either). If they don't work as well as expected, just add a few more features to take care of the errors, or find more suitable data. On the other hand, orthogonal expansions - Zernike moments, Fourier, Haar, Schwartzenegger - generate coefficients that are sensitive to any difference between shapes, whether essential or incidental. If we used them to recognize people, they would make a mistake every time they changed their hat. We need automatically generated features that are based on the observable difference between classes yet resist common, predictable sources of character distortion.

Enjoy your insight that optimal classifiers based on analytically tractable parametric distributions don't work very well except on artificially generated data, and that in hyperspace the boundaries between classes of live data aren't planes, paraboloids, or much of anything else that can be described by a neat equation. The realization that non-parametric classifiers with lots of adjustable parameters work better will surely follow. Then you can opt either for nearest neighbors and Parzen-windows based directly on the training samples, or for neural networks that internalize the sample vectors in a mere few thousand passes through the data. Whichever you pick, be assured that your initial algorithms for training *and* classification can be (and most likely already have been) improved upon. But please don't write about it.

Skeletons have many interesting quirks. Unmasking each quirk and then getting rid of it is an entire industry, but not the *OCR* industry. Although there is little evidence that *thinning* is a natural human activity or that alphabetic characters grow from the spine out, feature extraction and evaluation methods based on these ideas are popular. The hand-written U-V discrimination, first thoroughly studied twenty years ago, will keep the skeleton crews busy. Interestingly enough, in the sixties printed characters were sometimes *fattened.* Times change.

My list has many more items, but I don't want to sound grumpy. Let me mention instead some topics that turn me on.

5 New Ideas

Combining automated zoning, general-purpose layout analysis, font recognition and character classification leads to *page reconstruction.* Page reconstruction requires extracting sufficient information from a printed page image to generate the source code for generating a computer-editable facsimile version of the page. The target can be either a word-processor representation or a platform-independent page-layout language. The result can be searched, windowed, scaled, and manipulated like any computer-generated document. Although surprisingly good results have already been demonstrated, plenty of challenging problems remain. Among them are the identification of the best matches among the fonts available for reconstruction; imperceptible format perturbations to compensate for minor changes in font geometry; half-tone to dither conversion; table representation; and the generation of sensible code for formulas and equations.

One step beyond page reconstruction is *functional description* of arbitrary documents in a format such as SGML or ODA. While page reconstruction requires the identification of only physical attributes such as blank lines, indentations and page breaks, logical markup requires the recognition of functional components like subtitles, footnotes, paragraphs, and emphasis. Long term goals include partial automation of link-insertion. Most of the functional document image analysis reported to date stops short of converting the results to a widely usable format. This is not a trivial step. As already mentioned, the major impetus for functional description is the rapid growth of the World Wide Web, but conventional text retrieval will also benefit.

Complete systems of the type just mentioned cannot be developed by isolated researchers. Therefore the question of *intermediate representations* and data structures takes on additional importance. We hope that some of the current endeavors will lead to flexible, widely accepted representations for zoning, layout, typeface, segmentation, classification, context, and functional component identification for both test documents and reference data. Such representations must, of course, be compatible with current document interchange standards.

There is innovative research on *segmentation-free classification*, especially for hand-printed and cursive writing. But I would like to see an operational definition of the difference between segmentation-free classification and iterative or recursive segmentation-and-recognition. Much of this work is modeled on recent work on speech recognition. Linguistic constraints are often invoked, but I believe that there is hope for recognizing touching or overlapping characters even without them. I welcome the gradual erosion of the distinction between word and character recognition. These two processes surely work best in synergy.

Classifier combination remains an exciting topic. Voting methods have long been used for isolated character recognition, but robust string comparison algorithms have been only recently applied to align the output of "black box" OCR devices with the reference text for accurate counting of errors. It is possible that new OCR devices that provide word or character coordinate output will reduce the need for string matching, but this requires precise geometric registration of the page. For systems which also produce confidence measures, there are new methods based on rank-ordering. But will classifier combination survive if someone discovers how to combine disparate features into a single classifier?

The development of completely automated conversion systems for engineering drawings and maps is, in my view, a very long-term proposition. We should start with an interactive system such as AUTOCAD or CAD-OVERLAY that allows complete conversion of *any* drawing, instead of aiming at complete automation and relegating error correction to an afterthought. The human functions should gradually be taken over by the computer, starting with the easiest ones. The operator should use only AutoCad-type commands to enter parts of the drawing. Both the original bitmap and already converted portions can be shown as screen overlays. The parameters necessary for automating the process must be extracted from this type of user interaction only. For example, after the operator labels a resistor or two on the first drawing of a family, almost every resistor of the same size and shape should be automatically recognized. The converted documents should, of course, be in some standard format such as DXF.

More generally, it is time to concentrate on *systems that improve with use.* Now that even commercial OCR systems run entirely in software, nothing prevents continued fine-tuning of system parameters. We need more attempts to introduce feedback from down-stream programs (information retrieval, accounting, or schematic analysis) to alter layout analysis or classification parameters rather than simply correct individual errors. Current systems don't even exploit human post-editing in a manner analogous to adding "custom" words to ordinary spell-checkers. While these methods require only supervised adaptation, we should not rule out unsupervised learning. The basic premise is that the mass of data processed by any DIA or OCR systems in daily operation is more representative of future data than anything available at the factory.

6 Automated Evaluation

Automated benchmarking is not all that new. In the fifties and sixties, robot typewriters and printers pounded out hundreds of thousands of documents for testing MICR and OCR font readers. The IBM 1975 was tested on 1.3 million *pages.* Academic researchers reported results mainly on isolated characters. Although many different public test sets were produced and distributed through the Computer Society Repository, none attained the popularity of the hand-digitized data released by Highleyman in 1963. Competitive large-scale tests were conducted for address reading and, more recently, for census forms. Several large Japanese character sets were compiled. As the size of the test data sets grew, they migrated to CD-ROM. In the last few years, there has been renewed emphasis, and significant progress, on testing OCR and zoning accuracy on complete pages.

As has been painfully demonstrated over and over, large data sets don't guarantee accurate prediction of performance in the field. The data must also be *representative.* Sampling schemes reported to date have been based more on expediency than on sound statistics: we need to develop stratified and sequential sampling designs for large document populations. I am not aware of any organization or agency that has reported a well-planned *document census* centered on OCR and DIA variables.

In addition to the nature of the sample, we must consider the *sample size.* Even if the errors are statistically independent, it takes a sample size 1000 times greater to achieve a given level of significance at an error rate of 0.1% than at 1.0%. (Most researchers don't, however, report *any* confidence intervals). Underlying every statistical test there is some independence assumption that must be carefully examined.

Regardless of the specific performance aspects ("*metrics*") being evaluated, I find it convenient to divide methods of automated evaluation into five major paradigms. The first is based on *manually labeled samples*. The accuracy of the labels can be ensured by comparing the results of data entry by two independent operators in the spirit of key-punch verification. Isolated characters are identified by a serial or accession number, and errors can be readily counted and classified. For text data, sophisticated string-matching algorithms are used to align the OCR output with the reference data. Character errors, word errors, error-reject trade-offs, and zoning accuracy have all been reported on large, manually-labeled document data bases.

The second paradigm is based on the *cost of post-processing*. In some applications, including layout analysis and line-drawing conversion, there exists no truly satisfactory classification of errors. In these cases, the cost (or time) of correcting the errors manually may be used as a performance measure. Even for text, character errors do not reflect post-editing costs accurately. However, the components of the edit-distance that are a by-product of string matching may yield a more satisfactory approximation.

The third paradigm measures performance on *pseudo-random defect models* that generate bitmaps directly. Elaborate defect generators have been constructed for both isolated characters and printed pages. Such models are useful only to the extent that they mirror the distortions found in some real population of documents. The validation of defect models is the objective of vigorous research.

The fourth paradigm, based on *synthetic documents*, must be distinguished from the third. It too is based on bitmaps generated by a word-processor or layout language that mimic real documents. These bitmaps are, however, "ideal" bitmaps that are not perturbed by artificial noise. The reference data required to evaluate performance is the source code of the synthetic document, and inherently contains every typographic or layout detail that might be of interest. The noise component is modeled by printing, copying, and scanning the synthetic documents. Regardless of the faithfulness of the format and typeface reproduction, such synthetic documents are useful only to the extent that the printers, copiers and scanners used to produced them are representative of some actual application. This method seems restricted to printed documents and computer-generated maps and line-drawings. (However, hand-print classification research could take advantage of writing test samples with both real and "electronic" ink, and retaining the temporal trace of the stylus for comparison with off-line classification.)

The last paradigm, *goal-directed evaluation*, may eventually supersede all others. Except for digital library applications, OCR and DIA are not, after all, intended to produce output for human consumption. Converted documents will be processed by some downstream program for information retrieval, accounting, market analysis, or circuit simulation. Most errors can be detected, at least in principle, by these programs. Goal directed evaluation requires, however, the availability of downstream programs and large databases beyond the reach of most researchers.

Because there has been only very limited success in analytically modeling DIA and OCR systems, sound, objective and reproducible experimental evaluation is of the utmost importance. Therefore the current interest in large-scale automated evaluation of a range of performance measures is most welcome. In the United States, extensive DIA and OCR evaluation activities have been recently reported by CEDAR, ISRI, NIST, and the University of Washington.

7 Conclusion

Did I say that DIA and OCR may be approaching a mid-life crisis? Perhaps it is only puberty. They are surely on the threshold of success and prosperity. The best pickings still await us in this fertile patch at the intersection of pattern recognition, image analysis, and artificial intelligence.

Acknowledgment

The author gratefully acknowledges the influence of knowledgeable colleagues and of excellent students, past and present. Some of them commented kindly (and some less kindly) on drafts of this paper. He is also appreciative of the sustained support of the Central Research Laboratory, Hitachi, Ltd. and of the Northern-Telecom/BNR Education and Research Networking Program. This work has been conducted in the New York State Center for Advanced Technology (CAT) in Manufacturing, Automation and Robotics, which is partially funded by a block grant from the New York, State Science and Technology Foundation.