DOCUMENT IMAGE ANALYSIS: AUTOMATED PERFORMANCE EVALUATION

George Nagy
Rensselaer Polytechnic Institute
Troy, NY 12180, USA

ABSTRACT

Both users and developers of OCR systems benefit from objective performance evaluation and benchmarking. The need for improved tests has given additional impetus to research on evaluation methodology. We discuss some of the statistical and combinatorial principles underlying error estimation, propose a taxonomy for reference data, and review evaluation paradigms in current use. We provide pointers to recent work on the evaluation of isolated character classification, text reading, layout analysis, interpretation of hand-printed forms, and line-drawing conversion.

1. Introduction

In optical character recognition (OCR) and document image analysis (DIA), automated benchmarking is gradually reducing exclusive reliance on advertising claims and on informal product tests conducted by technical magazines. *Benchmarking* is defined here as the comparative evaluation of some specified characteristic of several products designed for the same purpose: *automated benchmarking* suggests objective, quantitative criteria based on large scale tests. The measure used for comparison (e.g., throughput, error rate, reject rate) is called a *metric*. More generally, OCR and DIA testing spans the gamut of scientific experimentation, detailed system assessment for improved product design, and marketing-oriented comparative product evaluation.

Automated benchmarking is not 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 in 1965 on 1.3 million pages of printed and typewritten social security earnings reports. 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. Researchers reported innumerable comparisons of feature sets and classifiers on isolated characters. In the seventies and eighties, large-scale tests were conducted for postal applications. Several digitized Japanese character sets were compiled and exercized. 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.

Another recent development is the systematic testing of "black-box" OCR systems by an independent organization instead of distributing the test data [ISRI 93, 94, 95].

A brief overview of benchmarking paradigms and of performance metrics for character accuracy, word accuracy, zoning, layout, document quality, and throughput, is presented in [Kanai 93]. Nartker discusses the basis of scientific measurement in information technologies and its application to the evaluation of document processing systems. He emphasizes the importance of better measures both for describing existing documents and for quantifying the results of processing [Nartker 94]. Recent test data bases and benchmarking procedures are described in detail in [Nartker 92; ISRI 93, 94, 95; Phillips 93a, Phillips 93b, Haralick 94, Wilkinson 92, Geist 94]. A report of the DAS'94 Working Group on Evaluation Criteria for OCR and Document Analysis Systems appears elsewhere in this volume. Here we examine the principles underlying evaluation methods, and sample what has been done and what remains to be done in specific areas of document analysis. We build on a point of view defined in an earlier paper:

> The customary evaluation paradigm requires comparing an *observed variable* with a *reference variable* (ground truth) under *controlled conditions*. The observed variable is typically some output of the document analysis system (such as character identities or paragraph markers), the reference variable is the desired output, and the controlled variables (conditions) are selected attributes of the input image over a sample document domain [Kanai 93].

There are many difficult choices underlying every aspect of this paradigm. Among them we single out for attention *sampling considerations*, and the type and source of the *reference data* to be used. A third important aspect, the choice of *observed variables*, depends on the specific application area, and will be discussed in separate sections for each area: printed characters, layout, handprinted characters, and schematic diagrams. We shall look at both established and evolving methods. Our further objective here is to interest the larger computer and decision science community in the design of accurate, reliable, and economical tests.

2. Sampling Considerations

Since we are concerned only with experimental evaluation, the first step is the choice of a sampling scheme. How can we ensure that the selected sample is *representative* of some application, i.e., that measurements on the sample will predict real-world performance? How large a sample is needed for reliable estimates? What are the pitfalls to be avoided?

OCR and DIA errors tend to exhibit far greater variation between documents than between competing systems. What is usually wanted is an estimate of the average of some performance parameter over the entire document set (*population*) from which the sample was drawn. It takes a much larger sample size N to estimate the parameter p of a Bernoulli variable to ten percent accuracy when $p = 0.01$ or 0.99 than when it is 0.5. If the error rate is low, i.e., p is near zero, then the confidence interval as a fraction of p is proportional to $N^{-1/2}p^{-3/2}$. To put it another way, for the same percentage uncertainty, we need 1000 times more samples for $p = 0.01$ than for $p = 0.1$. In such situations normality assumptions are suspect, and non-parametric methods of estimating statistical error, such as rank-order statistics, the bootstrap, the jack-knife, and cross-validation, may be more appropriate [Efron 83, Jain 87, Dudewicz 88, ISRI 95]. An excellent discussion of small sample effects in estimating classification errors is [Raudys 91].

How the sample set is selected is as important as the size of the sample. It might seem that only random selection can guarantee a representative sample. Consider, however, what happens if 10% of the pages are so bad that an OCR system will make about 25% error on them, while on the rest the errors are negligible. We will then obtain, on a small sample, the most accurate estimate of the population error rate using the appropriately weighted error rate on low-quality documents. The relevant sample size N here is the number of documents, not the number of characters.

In general, if the population is not homogenous (i.e., the sample variables have different variances in sub-populations), then stratified and sequential sampling designs are more efficient than random sampling [Brunk 68]. Sampling schemes reported to date have been based more on expediency than on sound statistical considerations. For the accurate estimation of the cost of large-scale document conversion, it will be essential to develop stratified sampling designs based on a *document census*.

Unless there are position-dependent errors due to printing or scanning, errors among characters from different documents can be assumed to be independent events whose sum obeys the binomial probability law. Under this assumption, Casey's Last Theorem states that for a reliable estimate, it is sufficient to commit 10 errors [Casey 1980]. If the population error rate is p, and N characters are required to obtain the first 10 errors, then the estimate $10/N$ of p will be off by more than 30% with a probability of only 5%. (This is not, however, an *unbiased estimate* of p, because it is based on the negative binomial distribution for N.)

OCR errors are usually clustered among confusion sets of similar characters such as {e,c,o} or {I,l,],/}. The confusion sets depend on typeface and type size [Pavlidis 83]. In many applications, punctuation errors are cannot be neglected. Error statistics obtained from alphabets must, of course, be weighted by normal letter frequencies (in text, easily-confused x-height characters are prevalent).

Correlated samples must be avoided: for example, it is safer to base estimates of the probability of e/c confusions on independently selected documents than on an equal number of errors on the same document. Correlated samples are all the more perfidious because independence cannot be verified by statistical tests. The distribution of the correlation coefficient under the null hypothesis is necessarily based on an independence assumption, therefore rejecting the null hypothesis can only confirm statistical *dependence.*

Systematic bias, for instance from a fixed binarization threshold, or from a poorly-designed scaling algorithm to change the spatial sampling rate, should also be avoided. An insidious cause of optimistic bias is re-use of the same test data for testing successive improvements of a recognition system [Nagy 83, Raudys 91].

3. Reference Variables

The observed performance metric, considered above as a random variable whose distribution is to be estimated, is the difference between the results obtained by DIA on the selected samples and the reference results on the same samples. We are aware of five general methods for conducting the comparison. The first three, based on manually labeling, artificial data, and synthetic documents, are explicit. The other two, based on the costs of post-processing and on measurements of the performance of downstream programs, are implicit.

3.1. Manually Labeled Test Data

Obtaining reference data, even with computer-aided methods, usually requires considerable human effort. The reference data consists of attributes, such as character identities, column boundaries, or circuit element categories and connectivity, extracted from the original document. Even in the case of simple character identities it is impossible to foresee every coding requirement. In the ISRI data set, for instance, all symbols outside the standard ASCII code are tagged with a tilde [Rice93]. The University of Washington CD-ROM uses TeX escape codes [Phillips 93]. Unicode encodes tens of thousands of writing symbols from all languages, but it does not attempt to represent graphemes [Bettels 93]. (Nowadays anybody can easily create entirely new graphemes using a font-design program.) Coding conventions for typefaces and layout are even more ambiguous, but ARPA's DIMUND project is attempting to promote a standard representation based on SGML [RAF 94].

It is clear that the reference data must be labeled at least an order of magnitude more accurately than the expected accuracy of the system to be tested. This is usually accomplished by double-keying and repeated reconciliation of the resulting reference files. OCR processing is sometimes used instead of one of the manual input streams. The low final residual error is very difficult to estimate accurately [Ha 94].

3.2. Artifical Test Data

With improvements in OCR classification algorithms (or, at least, in computer speed and memory size), it is becoming more and more expensive to obtain large-enough test sets. At the same time, digital fonts make it easier to generate character images with pseudo-random noise. For an example of a large-scale test on pseudo-randomly generated data, see [Baird 92]. While such artificial data is certainly adequate for comparing classification methods under controlled conditions, it can predict performance on real data only if the distortion generator reproduces the noise introduced in the printing, copying and scanning processes [Nagy 94].

Statistical testing of pseudo-random character generators is complicated by the fact that the noise processes that affect individual pixels cannot generally be considered independent [Sarkar 94]. So far there is little evidence of success in modeling the statistical dependencies necessary for accurate modeling of the distortions observed in scanned documents. The validation of pseudo-random generators is itself an active area of research [Baird 93, Li 94a, Li 94b, Nagy 94, Kanungo 94].

3.3. Synthetic Documents

Instead of perturbing bitmaps, test documents may be produced by mimicking the format of documents of interest using a page-layout language. The resulting synthetic documents can be printed, copied, scanned, and recognized [Esakov 94a, 94b]. Such synthetic documents are very different from documents produced by pseudo-random defect models. Synthetic documents are generated from an "ideal" bitmap, but they actually printed and scanned. The noise in such documents is therefore real printing and scanning noise, although it may not mirror the printing and scanning conditions found in the actual application. The degree of approximation can be estimated by comparing the results on the synthetic document and on the real document being imitated. Synthetic documents share with randomly-perturbed bitmaps the advantage of readily available, accurate, reference information.

In document analysis, the reference data may be in the form of a *synthetic digital reference document* encoded in the design-automation format to be produced by DIA. For instance, a printed page may be encoded using a layout language such as FrameMaker, text only may be encoded with a typesetting language such as TeX or troff, and engineering drawings may be converted to a CAD representation such as DXF. (PostScript is too low level to be useful as a DIA target language.) Although this takes much more effort than encoding only selected features, the resulting data structure reflects the purpose of document conversion more accurately. Note that here the actual, real-world document is scanned and recognized, and only the reference information is derived from the model.

3.4. Cost of Post-processing

An indirect approach to comparison with reference data is measuring the cost of the post-editing necessary to obtain some specified accuracy. This method cannot be automated. Furthermore, neither estimating the cost of post-processing, nor determining when the corrected text meets the performance specifications, is easy. Post-processing time varies widely with operator skill, and performance specifications are often loosely stated [Bradford 91a, Dickey 91]. An analysis of the cost of the efforts to obtain accurately encoded (in SGML) patent documents for the European Patent Office are reported, with many interesting observations, in [Stabler 94]. However, once the documents have been processed and edited, they provide a useful test data base for the automated evaluation of improved DIA algorithms. Measuring the cost of post-processing is particularly attractive in line-drawing conversion, where errors are difficult to classify and measure, and where it is difficult to synthesize accurate replicas of hand-drawn diagrams [Yu 94].

3.5. Goal-directed Evaluation

The fifth paradigm is to consider the effect of errors on downstream programs. Except for electronic library applications, documents are not converted to computer-readable form for the benefit of human readers. Examples of downstream processes that may catch DIA or OCR errors are book-keeping and accounting software, mail sorters, information retrieval programs, and circuit analysis packages. In view of the rather arbitrary nature of most OCR specifications, measuring downstream effects ("number of letters delivered to the wrong city") may be a more sensible approach. (However, DIA may be only one of several sources of error in such processes.) The rationale behind goal-directed evaluation is discussed further in [Trier 95].

This completes our discussion of reference data, and we are now ready to consider the observed variables for specific performance attributes. Because these attributes tend to be highly application-dependent, the remainder of the paper is organized according to the type of material being processed.

4. Printed Characters

Although the recognition of individual character shapes has spawned many interesting statistical and structural classification methods and a large body of literature, it is only tenuously related to OCR performance on printed text. In the latter, accurate location of paragraph, line, word and character boundaries, and the use of morphological, lexical, and syntactic context, are often the dominant factors. Because the evaluation of isolated pattern recognition is much simpler than that of OCR readers (a consideration that may account for the popularity of shape classification in academic research), we discuss them under separate headings.

4.1. Isolated Characters

Classification algorithms for isolated characters are usually evaluated on printed digits, upper or lower case alphabets or, occasionally, a larger symbol set including punctuation and special characters, or Chinese ideographs. The data may include several typefaces and sizes, but typically only one or two printers and a single scanner are used. (Before the advent of desktop printers, OCR manufacturers used "robot" typewriters and systematically varied ribbon and paper quality.) To obtain a range of qualities, copies of various generations may be used. The data is labeled by its position on the page. Under these circumstances, it is simple to count the number of errors produced by the recognizer.

Difficult discriminations are often highlighted in a confusion matrix [Kanai 93]. Since the cost of undetected substitution errors versus that of rejected characters varies from application to application, the entire trade-off is often plotted as an error-reject curve [Chow 70, 94]. This curve may be obtained from the confidence levels associated with each decision [Geist 94].

A recent example of isolated character classification, using neural networks, appears in an article by researchers from Stanford University and Canon Research Center America. They report *zero* errors on each of two independent tests on 1,047,452 Courier characters and 347,712 characters in twelve typefaces and type sizes ranging from 8 to 12 points. The alphabet consists of 94 symbols. Misclassifications resulting from "segmentation error, scanning error, and paper residue" and confusions between different characters that look identical in different fonts, are not counted as recognition errors and are not reported [Avi-Itzhak 95].

4.2. Text

The major sources of error on formatted text include segmentation failures in locating text lines, and word and character boundaries. (Word gap location may seem simple to the uninitiated, but tests indicate that missed and erroneously introduced blanks are very common [ISRI 94]). Failure to consider these factors has resulted in many overly optimistic predictions [Schantz 82].

OCR errors on text may be determined by (1) proofreading, (2) comparison between reference data and OCR output at a given coordinate location on the page, or (3) string comparison. The first method is not dependable for large volumes of text. The second method depends both on accurate registration of the page image with the reference image, and on recording the location of each item classified by the OCR device. Character coordinates are provided so far by only a few commercial devices. The third method is the only one currently used in large-scale tests.

String comparison based on dynamic programming to find the edit distance is relatively new to OCR, but it is an established and active topic in computer science [Wagner 74, Shankoff 83, Srihari 85, Bunke 92, Kukich 92]. Applications of string comparison to multiple-classifier voting were reported in [Bradford 91b, Handley 91]. For benchmarking, the strings of character labels emitted by an OCR system are compared with reference strings, and the differences are considered errors. Each printed line [Esakov 94a,b; Haralick 94] or each block of text [Rice 92, 93] may be considered as a single string. The alignment of the strings is affected by character segmentation errors, so a robust alignment algorithm is required.

An appropriate measure of the difference between two strings is the (minimum) edit distance, which depends, in turn, on the weights assigned to insertions, deletions, and substitutions [Shankoff 83, Rice 92, Esakov 94a, Esakov 94b, Sandberg 93, Geist 94]. For instance, one may consider a single-character substitution error equivalent in cost to that of an insertion (of a missed character) plus that of a deletion (of an extra character), or equivalent to just one of these. With any assignment of weights, there may be several combinations of edit operations that yield the same distance. The number of "segmentation" errors cannot always be distinguished, even in principle, from the number of "substitution" errors (consider, for instance, "mn" misinterpreted as "nm").

The weights required to find the edit distance are usually defined over single symbols only, but Esakov, Rice and Sandberg consider assigning different weights to multiple errors. Taking into account dependent errors could conceivably be extended to model word, phrase, line, or sentence accuracy. Currently, various types of word errors, including function words, stop words, key-words, and proper names, which may be relevant for specific applications, are obtained directly from the strings aligned to minimize character edits.

The edit distance itself is not necessarily a good measure of OCR correction costs. There is no guarantee that a human post-editor would actually execute the optimal edit operations. However, the difference strings found above as a byproduct of string comparison may be analyzed to assign penalties that reflect more closely the cost of various types of multi-character errors [Sandberg 93]. Scattered errors are generally considered more costly than clustered errors.

All of the methods described above yield an error rate for every zone or every page. How does one combine the error rates for samples of many zones or many pages? One may either take an unweighted combination of the error rate on each page, or count the total number of errors and divide it by the total number of words or characters. The latter seems more sensible, but the former allows every page error rate to be considered an independent sample, which facilitates setting confidence intervals for the overall error [ISRI 95].

5. Layout

Layout analysis includes *preprocessing*, *zoning* (text/non-text separation), *page recomposition* (the recognition of format and typography), and *functional labeling* of document components. With regard to the latter, we distinguish between variable-format and constrained-format applications, though the distinction is more a matter of degree than of kind.

5.1. Preprocessing

The major preprocessing functions are binarization, noise removal, and deskewing. A good deal of effort has been directed at the evaluation of thinning and skeletonization algorithms, but since thinned patterns are not used as input to any off-the-shelf DIA or OCR system, we will not consider them here. For the same reason, we omit feature evaluation research. Noise filters could be used after scanning, but we are not aware of any published evaluations of noise filters on live documents.

Earlier thresholding and edge-detection algorithms for postal addresses were compared in [Palumbo 1985]. Two recent evaluations are based on the error rate of commercial OCR devices run at every possible threshold setting [Smith 93, Grover 94]. They show conclusively that published algorithms set global thresholds that are far from optimal for any OCR device. Furthermore, the optimum threshold depends on the specific OCR system tested. Trier et al. compare eleven local and four global algorithms in terms of error and reject rates on hand-printed soundings from hydrographic charts, using an trainable experimental OCR module [Trier 95]. Although these studies clearly differentiate the available algorithms, we expect that thresholding will diminish in importance as OCR vendors shift to gray-scale processing.

Skew estimation algorithms are compared either to visual estimates, or tested on artificially rotated images. It appears that on plain text of more than more than a few lines, skew detection algorithms are competitive with human skew estimation. Residual skew of less than half a degree does not seem to affect OCR per se, but may hamper line finding. Accurate skew correction is particularly important for ruled forms and tables.

5.2. Zoning

There are at least three direct methods to evaluate the accuracy of zoning algorithms that locate the text on the page and discard all other material.

The first method tests certain aspects of zoning accuracy by string comparison based on the reading order of the characters on the page. This method does not depend on the data representation used for the zones, but it does require an OCR device that processes automatically zoned versions of the pages. The resulting character strings are compared to the output of the same OCR device run on correctly (manually) zoned versions of the same pages. Not only the edit distance, but the actual edit operations must be determined. OCR errors are subtracted from the zoning errors, which typically consist of long contiguous strings to be moved, inserted, or deleted. The comparison algorithm differs from the one used for OCR because swaps (transpositions) must also be located [Kanai 95].

The second method consists of comparing the automatically determined zone boundaries with the minimum and maximum boundaries of the manually obtained reference zones [Haralick 93]. Here the zone coordinates must be externally available, which is not generally the case with commercial OCR devices.

The third method also requires manual zoning but, instead of comparing the boundaries of the automatically and manually obtained zones, it compares the overlap of their respective foreground (black) pixel sets. Quantitative measures of zone quality are derived from this information. This method has been extensively tested and provides sufficient detail to be used for improvement of automated zoning algorithms [Randriamasy 94].

Indirect evidence of zoning problems shows up as a byproduct in information-retrieval based tests. If some of the graphics areas are erroneously identified as text, the OCR system attempts to recognize graphic segments such as dashed lines as alphabetic symbols. Such graphic debris results in a large number of mismatches when the output of the OCR system is checked against a lexicon [Taghva 94].

5.3. Page Recomposition

In addition to obtaining computer access to the textual contents of a printed document, it may be necessary to recreate it in a form close to the original. In this case, the layout, typography, and illustrations must all be preserved. Merely displaying unformatted ASCII would deprive readers of essential visual clues about the organization of the material, while a facsimile reproduction as a bitmap image would be difficult to search for keywords. (But for methods to search bitmaps for words, see [Ho 91, Hull 93, Khoubyari 93, Bagley 94]). One solution to this dilemma is the generation of an accurate description of the printed document in a word-processor format or in a layout language like FrameMaker or Publisher. (Another solution is dual representation: ASCII for computer search, and facsimile bitmap for human readers).

Page recomposition is an important current goal in OCR. A good page recomposition facility should preserve typeface, type size, style attributes such as bold, italic, underlining, superscripts and subscripts, spacing (margins, centering, justification), and even row-column relations of tables.

Naturally, the accuracy of the format description must be evaluated. It is difficult to extract and manually record all the essential layout and typographic attributes of a document. Consequently, evaluation based on synthetic documents may be more economical than comparison with a manually generated attribute database.

As mentioned earlier, synthetic documents may be generated from a paper document by scanning the original, passing it through an OCR device, manually editing the result, and adding the necessary formatting information using a layout language. The result is a symbolic document that mimics the appearance of the original. It can be printed, scanned, and processed by the layout recognition software [Viswanathan 89]. The output of the layout recognition program can be translated into the source language used to (re)create the document in the first place, and compared to the code that was used to generate the synthetic document. We call this method *inverse formatting* [Kanai 93].

Unfortunately (from the evaluator's perspective), the relationship between layout source code and the resulting digital document is many to one. For instance, six 12-pt blank lines may be generated by a command to skip 6 lines at the current line spacing, or by a command to skip one inch. It is, however, possible to print the document from the layout code generated by the recognition software and visually compare it to the synthetic document [Okamoto 92]. In principle, it would be possible to compare the bitmap of the recreated document with that of the original, but a very sophisticated (elastic) registration method would be required. The problem is analogous to that of evaluating lossy image compression algorithms, and clearly requires further research.

5.4. Functional Labeling

In information-retrieval related applications, it is desirable to label each major document component prior to subsequent processing. In a journal article, important components might include the title, byline, authors' affiliation, copyright notice, abstract, illustration, and so forth [Krishnamoorthy 93]. In a business letter, the sender, addressee, date, and reference might have to be identified [Hoch 93, Dengel 94]. Several systems for identifying the language (English or French, Chinese or Japanese) of the document (or part of the document) have also been demonstrated [Ittner 93, Spitz 94]. Although dozens or even hundreds of documents may be tested, we are not aware of any attempts at *automated* evaluation.

In principle, it would be possible to compare the results of labeling to manually labeled data. However, the preparation of a reasonable-sized database is too expensive for most research projects. There is no current agreement regarding the data structures to be used. Both ODA and SGML are flexible enough for the purpose, but require agreement on the specific attributes to be tagged.

It would appear that here too large-scale automated evaluation must eventually be based on downstream processing. If, for instance, the authors of an article, the affiliation, or the sender, cannot be found in a list of authors, known affiliations, or prior correspondents, then the identification must be deemed erroneous. In language identification, the words can be checked against a lexicon. However, such goal-directed methods require flexible omnifont OCR capability. Because research organizations tend to develop only a part of the overall system, they seldom have access to the necessary verification tools.

5.5. Constrained Formats

Postal address block location and address reading. This is a highly specialized application that has been the subject of intensive research for over thirty years [Hessinger 68, Genchi 68]. Evaluation criteria used to depend on whether incoming (9-digit ZIP code, street address) or outgoing (5-digit ZIP code, destination country, state or city) was sorted. Current machines attempt to read the entire address, and encode it with a sprayed-on 11-digit bar code that represents the delivery order on the mail-carrier route. Different criteria apply to first-class mail, flats, and packages, and speed is an important consideration. The task is aided by a 300 megabyte country-wide directory of valid addresses that is updated weekly. A good starting reference to the voluminous literature on the subject is [Srihari 92], which includes some performance figures. A carefully constructed postal database collected from live mail is available for testing algorithms [Hull 94].

Dictionaries, directories, catalogues and bibliographies. The common denominator here is the presence of many short entries in essentially the same format. The entries have both mandatory and optional components, which are often distinguished by typesetting conventions. Some erroneously-processed entries can be rejected if mandatory components are missing. Others may be flagged using an application-specific database. For instance, prices or dates may be outside some acceptable range, certain dictionary tags can belong only to a few types, and strict syntactic conventions govern the naming of organic compounds. Some directories are updates of older versions which, if already available in electronic form, may be used to screen the output. We are not aware of any general approach that would allow use of existing ancillary information to check the accuracy of the output, but since these documents may be viewed as homogenous collections of essentially similar components, moderate sized samples should be adequate.

Tables. Most commercial OCR devices convert tables to lines of text. Even if that is acceptable, ISRI tests show that pages containing tables have many more errors due to automated zoning than other pages [Kanai 95]. But the real challenge with tables is to associate each entry with its proper row and column header for content-based retrieval from a relational database or a spreadsheet [Watanabe 91, Laurentini 92, Krishnamoorthy 94]. This is particularly difficult with multiply nested tables. For evaluation it would be useful to have a large set of tables with the corresponding information already available in a database or spreadsheet.

Formulas. While the appropriate output format is fairly clear for directories and tables, it is anything but obvious for mathematical formulas or equations. One possibility is inverse formatting, as described above, that allows manipulation of the symbols in some formula-setting language [Chou 89, Okamoto 92]. It would be, however, far more desirable to preserve the semantic information in a form usable by symbolic math programs.

Business letters. Depending on the diversity of the source, letters can be considered either constrained or unconstrained with regard to format. Parts of a letter may have a fairly fixed structure that can be analyzed with an attributed context-free grammar, while the main body requires natural-language processing techniques. For partial evaluation, on a small sample, of the different components of a complete letter-routing system, see [Dengel 94]. Orders and invoices also straddle the line. Within a single organization are normally produced on the same form or by the same software. But from the point of view of the recipient, who must process incoming mail from many sources, they are highly variable.

Miscellany: calling cards, resumes, classified ads, musical scores. Systems for the analysis of such material have been demonstrated, but we are not aware of any attempt at automated evaluation.

6. Handprinted Characters

Hand printing is notorious for its variability depending not only on the writers, but also on the particular task set for collecting the writing samples [Bakis 68]. Evaluation is necessary both within writers and across writers, and very large samples are necessary for stable results.

IEEE Repository. Much of the early research on handprinted character recognition made use of the "pattern recognition data bases" of the IEEE Computer Group Repository, established in the late sixties. The May 1974 issue of *Computer* magazine lists seven databases, ranging from 500 to 100,000 characters, including hand-print, print, and script. Hundreds of experiments were reported on a set of 2300 handprinted and hand-digitized characters on punched cards [Highleyman 63].

NIST. The National Institute of Standards and Technology has recently conducted two large scale tests with participants from dozens of academic, non-profit, and commercial organizations. The first test offered unsegmented characters without any sensible context: the evaluation was based on error-reject curves. The training data, distributed beforehand, was subsequently shown, by means of cross-validation statistics, to exhibit a markedly different distribution from the test set [Wilkinson 92]. The second test contained data from three different fields of the 1980 census forms. In this test two different scoring methods were used. The first simply counted the number of fields that were misclassified, without taking into account the degree of misclassification, i.e., the number and position of misclassified characters. The second method was based on the Levenshtein distance, and counted the number of insertions, deletions, and substitutions when the recognized string was optimally aligned with the reference string. Both sets of test procedures, the scoring methods, the results, and some of the counterintuitive results produced by the evaluation methods, are thoroughly analyzed in [Geist 94]. Several test databases, containing hundreds of thousands of characters - some isolated, some on tax forms - and the evaluation software are available from NIST on CD-ROM [Garris 93, NIST 94].

Loral. A division of Loral Federal Systems (formerly part of IBM) has adopted the NIST scoring package to conduct independent evaluations, based on a customer's own data, on various commercial OCR systems. The company also offers facilities for image enhancement, field extraction, and contextual post-processing [Probst 94].

Bank checks. The courtesy amount field on bank checks offers several possibilities for validation. First, the courtesy amount may be compared to the legal amount (which, however, on personal checks is usually in script and much harder to read than the numerals in the courtesy field). Second, when the checks are deposited, the individual amounts may be compared to those listed on the deposit slip (which are not necessarily in the same order) [Chang 77]. Third, most remittances include an invoice return slip, usually in a machine readable font. Because of the high costs associated with errors, OCR is normally used to replace only one of the two operators in double-keying the data. While all of this information is used in routine processing, we are not aware of any published evaluations. Banks are understandably reluctant to release check images for research. For a discussion of current practices in financial data capture, see [Schantz 94].

Forms. Form imaging is a burgeoning field with many commercial products, but relatively little research has been published on the evaluation of form-segmentation systems. Most of the published results appear to have been obtained by visual inspection [Leibowitz 92]. In principle, the contents of the cells could be checked against appropriate ASCII databases, and the residual line-art could be compared for integrity to stored form representations.

7. Schematic Diagrams

Much effort has been devoted to the analysis of electrical and mechanical drawings, organization charts, and wiring diagrams [Kasturi 92]. However, large-volume automated evaluation still seems sufficiently far away to have been given little thought. So far, relatively few projects have combined sophisticated processing of line drawings with state-of-the-art OCR.

Most research efforts do not even convert the results of the analysis to the data structures produced when such drawings are prepared on CAD systems. The format produced by AutoCad, DXF, has become a de facto standard for mechanical systems, while most circuit-capture programs produce a Spice-compatible net-list. As in the case of layout languages, the mapping is many-to-one, rendering direct comparison impractical. Nevertheless, converting the drawings to a standard computer format is a necessary first step towards automated evaluation using downstream programs.

Some researchers evaluate their results by keeping track of the amount of human post-editing necessary to correct processing errors. Either the total time, or the number of editing operations are counted [Yu 94]. Here the same operator (usually the researcher) is responsible both for correcting the errors and for checking if the result meets specifications.

Among the highest volume applications are utility maps such as telephone and power cabling and piping diagrams. These applications are particularly appropriate for automated processing because of the long useful life of such diagrams: buried conduits and pipes remain in place for decades, and quick access to such information for updates and emergencies is a vital concern. Merging data from old blueprints with that currently produced by CAD systems is a major concern for telephone, electricity, and gas companies. It is, however, difficult for academic researchers to gain access to the relevant non-graphic information (stored in secure company databases) that would allow automated evaluation.

Cartographic data also has a long lifetime. While federal mapping agencies (USGS, DMA) update topographic and hydrographic maps every ten or twenty years through remote sensing and field surveys, much smaller governmental units are usually responsible for cadastral (property) maps. Hand-drawn maps entered into computer databases are typically thoroughly inspected and corrected by expert municipal employees [Boatto 92]. Releasing the digital data corresponding to existing and already converted and inspected maps would be of great help in developing automated evaluation methods, but such data is often considered confidential or privileged.

8. Conclusion

Improved computing technology has rendered it possible to process volumes of data large enough to render evaluation by inspection inpracticable. However, except for optical character recognition, where classical string comparison methods have been applied to significant volumes of data, little has been attempted by way of automated evaluation of DIA systems.

Some large text-oriented document databases with reference data are now available, but further research on statistically sound evaluation methods using such databases is necessary. Furthermore, the relationship of existing databases to actual DIA and OCR tasks requires serious attention. Without more purposeful and systematic document sampling techniques, results on such databases cannot predict actual application costs, although they can be used (cautiously!) for comparative studies.

In many applications, where an increasing amount of information must be extracted from the documents, the preparation of synthetic versions of existing paper documents in a layout language or CAD format is more appropriate than databases of extracted attributes. Such documents should be printed, copied, and scanned in a manner reflecting the actual application as closely as possible.

Currently available distortion and noise models may be adequate for internal algorithm development, but have not yet been shown to be sufficiently reliable for performance evaluation and prediction.

Since paper documents are seldom converted to computer form for human use, downstream software and databases may see increasing direct use in goal-directed evaluation.

Because the characteristics of documents are the result of centuries of evolution rather than methodical design, DIA and OCR researchers often have to resort to *ad hoc* methods with few formal or mathematical foundations. The evaluation of such systems may, however, lend itself to a more rigorous approach. We hope that this chapter will call attention to some of the exciting unsolved problems of DIA and OCR evaluation.

9. Acknowledgment

The author thanks Richard Casey, Rainer Hoch, Junichi Kanai, Steven Rice, and Tom Nartker for critical comments on early drafts. He also gratefully acknowledges the 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. The CAT is partially funded by a block grant from the New York State Science and Technology Foundation.

10. References

[Avi-Itzhak 95] *High accuracy optical character recognition*, H.I. Avi-Itzhak, T.A. Diep, H. Garland, IEEE Trans. PAMI 17, 2, pp. 218-224, February 1995.

[Bagley 94] *Editing images of text,* S.C. Bagley and G.E. Kopec, C. ACM 37, 12, pp. 63-72, December 1994.

[Baird 92] H.S. Baird, R. Fossey, *A 100-font classifier,* ICDAR-91, St. Malo, pp. 332-340, 1991.

[Bakis 68] R. Bakis, N. Herbst, G. Nagy, *An experimental study of machine recognition of hand-printed numerals,* IEEE Trans. SSC-4, 2, pp. 119-132, 1968.

[Bettels 93] J. Bettels, F.A. Bishop, *Unicode: a universal character code*, Digital Technical Journal 5, 3, pp. 21-31, Summer 1993.

[Boatto 92] L. Boatto et al., *An interpretation system for land register maps*, Computer 25,7, pp. 25-33, July 1992.

[Bradford 91a] R. Bradford, *Technical factors in the creation of large full-text databases*, preprint, DOE Infotech Conference, Oak Ridge, TN May 1991.

[Bradford 91b] R. Bradford and T. Nartker, *Error correlation in contemporary OCR systems,* Proc. Int'l Conf. on Document Analysis and Recognition, St. Malo, France, pp. 516-523, September 1991.

[Brunk 60] H.D. Brunk, *An introduction to mathematical statistics*, Ginn and Company, Boston, 1960.

[Bunke 92] H. Bunke, *Recent advances in string matching, in Advances in Structural and Syntactic Pattern Recognition*, (H. Bunke, editor), World Scientific, pp. 3-21, 1992.

[Casey 80] R.G. Casey, *The ρ-10 ("rotten") estimator*, personal cummunication, c. 1980.

[Chang 77] S.K. Chang and G. Nagy, *Deposit-slip-first check reading*, IEEE Transactions on Systems, Man, and Cybernetics 7, 1, pp. 64-68, January 1977.

[Chou 89] P. Chou, *Recognition of equations using a two-dimensional stochastic context-free grammar*, SPIE Conf. on Visual Communications and Image Processing, Philadelphia, PA Nov 1989.

[Chow 70] C.K. Chow, *On optimum recognition error and reject tradeoff*, IEEE Trans. IT-16, 1, pp. 41-46, Jan. 1970.

[Chow 94] C.K. Chow, *Recognition error and reject trade-off*, Proc. Third Annual Symposium on Document Analysis and Information Retrieval,, ISRI, Las Vegas, NV, pp.1-8, April 1994.

[Croft 94] W.B. Croft, S.M. Harding, K. Taghva, J. Borsack, *An evaluation of information retrieval accuracy with simulated OCR output,* Proc. Third Annual Symposium on Document Analysis and Information Retrieval,, ISRI, Las Vegas, NV, pp.115-126, April 1994.

[Dengel 94] A. Dengel, R. Bleisinger, F. Fein, F. Hones, M. Malburg, *OfficeMAID - A system for office mail analysis, interpretation, and delivery*, Procs. DAS '94, pp. 253-176, Kaiserslautern, 1994.

[Dickey 91] L.A. Dickey, *Operational factors in the creation of large full-text databases*, preprint, DOE Infotech Conference, Oak Ridge, TN May 1991.

[Dudewicz 88] E.J. Dudewicz and S.N. Mishra, *Modern Mathematical Statistics*, John Wiley and Sons, New York 1988.

[Efron83}, B. Efron and G. Gong, *A leisurely look at the bookstrap, the jackknife, and cross-validation*, The Americal Statistician 37, 1, pp. 36-48, February 1983.

[Esakov 94a] J. Esakov, D.P. Lopresti, J.S. Sandberg, *Classification and distribution of optical character recognition errors*, Procs. Symposium on Document Analysis, SPIE Volume 2181, pp. 204-216, February 1994.

[Esakov 94b] J. Esakov, D.P. Lopresti, J.S. Sandberg, J. Zhou, *Issues in automatic OCR error classification*, Proc. Third Annual Symposium on Document Analysis and Information Retrieval, ISRI, Las Vegas, NV, pp.401-412, April 1994.

[Garris 93] M.D. Garris, *NIST scoring package certification procedures*, NISTIR 5173, NIST, Advanced Systems Division Image Recognition Group, Gaithersburg, MD April 1993.

[Geist 94] G. Geist et al., *The second census optical character recognition systems conference*, NISTIR 5452, NIST, Advanced Systems Division Image Recognition Group, Gaithersburg, MD May 1994.

[Genchi 68] H. Genchi, K.I. Mori, S. Watanabe, and S. Katsuragi, *Recognition of handwritten numeral characters for automatic letter sorting*, Proc. IEEE 56, 1968, pp. 1292-1301.

[Grover 94] K. Grover and J. Kanai, *Evaluation of thresholding algorithms for OCR*, ISRI Technical Memo, UNLV, Las Vegas, NV, 1994.

[Ha 94] J. Ha, R.M. Haralick, S. Chen, I.T. Phillips, *Estimating errors in document databases*, Proc. Third Annual Symposium on Document Analysis and Information Retrieval, ISRI, Las Vegas, NV, pp.435-460, April 1994.

[Handley 91] J.C. Handley and T.B. Hickey, *Merging optical character output recognition for improved accuracy*, Proc. RIAO 91 Conference, Barcelona, Spain, pp. 160-174, April 1991.

[Haralick 93] R.M. Haralick, *English document database design and implementation methodology*, Proc. Second Annual Symposium on Document Analysis and Information Retrieval, ISRI, Las Vegas, NV, pp. 65-104, April 1993.

[Hennis 68] R.B. Hennis, M.R. Bartz, D.R. Andrews, A.J. Atrubin, K.C. Hu, *The IBM 1975 Optical page reader*, IBM J. Res. & Dev. 12, 1968, 345-371.

[Hessinger 68] R.W. Hessinger, *Optical character recognition in the Post Office*, in Pattern Recognition (L.N. Kanal, editor), Thompson,Washington, 1968.

[Highleyman 63] W.H. Highleyman, *Data for character recognition studies*, IEEE Trans. EC-12, pp. 135-136, March 1963.

[Ho 91] T.K. Ho, J.J. Hull, and S.N. Srihari, *Word recognition with multi-level contextual knowledge*, Proc. Int'l Conf. on Document Analysis and Recognition, St. Malo, France, pp. 905-916, September 1991.

[Hoch 93] R. Hoch and A. Dengel, *INFOCLA: Classifying the message in printed business letters*, Proc. Second Annual Symposium on Document Analysis and Information Retrieval, ISRI, Las Vegas, NV, pp. 443-456, April 1993.

[Hull 93] J.J. Hull and Y. Li, *Word recognition result interpretation using the vector space model for information retrieval*, Proc. Second Annual Symposium on Document Analysis and Information Retrieval, ISRI, Las Vegas, NV, pp. 147-156, April 1993.

[Hull 94] J. J. Hull, *A database for handwritten text research*, IEEE Trans. PAMI-16, pp. 550-554, May 1994.

[ISRI 93] *UNLV Information Science Research Institute, Annual Report*, 1993.

[ISRI 94] *UNLV Information Science Research Institute, Annual Report*, 1994.

[ISRI 95] *UNLV Information Science Research Institute, Annual Report*, 1995.

[Ittner 93] D.J. Ittner, H.S. Baird, *Language-free layout analysis*, Procs. ICDAR-93, pp 336-340, Tsukuba Science City, 1993.

[Jain 87] A.K. Jain, R.C. Dubes, C. Chen, *Bootstrap techniques for error estimation*, IEEE Trans. PAMI-9,5, pp. 628-633, September 1987.

[Kanai 93] J. Kanai, S. Rice, G. Nagy, T. Nartker, *Performance metrics for printed document understanding systems*, Procs. Inter. Conf. on Document Analysis and Recognition, ICDAR-93, Tsukuba, Japan, pp. 424-427, October 1993.

[Kanai 95] J. Kanai, S.V. Rice, T.A. Nartker, G. Nagy, *Automated evaluation of OCR zoning*, IEEE Trans. PAMI-17, 1, pp. 86-90, January 1995.

[Kanungo 94] T. Kanungo, R.M. Haralick, H.S. Baird, W. Stuetzle, D. Madigan, *Document degradation models: Parameter estimation and model validation*, Procs. IAPR Workshop on Machine Vision Application (MVA '94), Kawasaki, pp. 552-557, December 1994.

[Kasturi 92] R. Kasturi and L. O'Gorman, *Document Image Analysis: A bibliography*, Machine Vision and Applications 5, 3, pp. 231-243, Summer 1992.

[Khoubyari 93] S. Khoubyari, J.J. Hull, *Keyword location in noisy document images*, Proc. Second Annual Symposium on Document Analysis and Information Retrieval, ISRI, Las Vegas, NV, pp. 217-232,, April 1993.

[Krishnamoorthy 93] M. Krishnamoorthy, G. Nagy, S. Seth, M. Viswanathan, *Syntactic segmentation and labeling of digitized pages from technical journals*, IEEE Trans. PAMI-15, 7, pp. 737-747, July 1993.

[Krishnamoorthy 94] M.S. Krishnamoorthy and E. Green, *Recognition of tables using table grammars*, Procs. Proc. Fourth Annual Symposium on Document Analysis and Information Retrieval, ISRI, Las Vegas, NV, April 1995.

[Kukich 92] K. Kukich, *Techniques for automatically correcting word in text*, ACM Computing Surveys 24, 4, pp. 377-440, December 1992.

[Laurentini 92] A. Laurentini and P. Viada, *Identifying and understanding tabular material in compound documents*, Procs. ICPR-11, The Hague, pp. 405-409, 1992.

[Leibowitz 92] S. Liebowitz Taylor, R. Fritzson, J.A. Pastor, *Extraction of data from preprinted forms*, Machine Vision and Applications 5, 3, pp. 211-222, Summer 1992.

[Li 94a] Y. Li, D. Lopresti, A. Tomkins, *Validation of document image defect models for optical character recognition*, Procs. SDAIR-94, Las Vegas, pp. 137-150, 1994.

[Li 94b] Y. Li, D. Lopresti, G. Nagy, A. Tomkins, *Validation of image defect models for optical character recognition*, Technical Report MITL-TR-69-93R, Matsushita Information Technology Laboratory, Princeton, 1994.

[Lopresti 94] D. Lopresti, J. Zhou, *Using consensus sequence voting to correct OCR errors*, Procs. DAS94, pp. 191-202, Kaiserslautern, October 1994.

[Nagy 83] G. Nagy, *Candide's practical principles of experimental pattern recognition*, IEEE Trans. PAMI-5, 2, pp. 199-200, March 1983.

[Nagy 94] G. Nagy, *Validation of OCR datasets*, Proc. SDAIR-94, Annual Symp. on Doc. Analysis and Retrieval, pp. 127-136, Las Vegas, April 1994.

[Nartker 92] T.A. Nartker, R.B. Bradford, B.A. Cerny, *A preliminary report on UNLV/GT1: A database for ground-truth testing in document analysis and character recognition*, Proc. First Symposium on Document Analysis and Information Retrieval, Las Vegas, NV March 1992.

[Nartker 94] T.A. Nartker, *Need for information metrics: with examples from document analysis*, Procs. Symposium on Document Analysis, SPIE Volume 2181, pp. 184-193, February 1994.

[NIST 94] *Special Databases 1-9*, NIST, Advanced Systems Division Image Recognition Group, Gaithersburg, MD May 1994.

[Okamoto 92] M. Okamoto and A. Miyazawa, *An experimental implementation of document recognition system for papers containing mathematical expressions*, Structured Document Image Analysis, (H.S. Baird, H. Bunke, K. Yamamoto, editors), Springer-Verlag, 36-53, 1992.

[Palumbo 86] P.W. Palumbo, P. Swaminathan, S.N. Srihari, *Document image binarization: evaluation of algorithms*, Procs. SPIE Symp. on Digital Image Processing, San Diego, 1986.

[Pavlidis 83] T. Pavlidis, *Effects of distortion on the recognition rate of a structural OCR system*, Procs. CVPR-83, Washington, June 1983, pp. 303-309.

[Phillips 93a] I.T. Phillips, S. Chen, R.M. Haralick, *CD-ROM document database standard*, Procs. ICDAR-93, pp 478-484, Tsukuba Science City, 1993.

[Phillips 93b] I.T. Phillips, J. Ha, R.M. Haralick, D. Dori, *The implementation methodology for a CD-ROM English document database*, Procs. ICDAR-93, pp. 478-484, Tsukuba Science City, 1993.

[Probst 94] R.E. Probst, W.W. Klein, G. Meyers, *Accuracy considerations in employing optical character recognition for automated data capture*, Today 17, 1 (The Association for Work Process Improvement), pp. 10-16, September 1994.

[RAF 94] RAF Technology, *Document Attribute Format Specifications*, Technical Report, RAF Technology Inc., Redmond, WA October 1994.

[Randriamasy 94] S. Randriamasy, L. Vincent, *A region-based system for the automatic evaluation of page segmentation algorithms*, Procs. DAS94, Kaiserslautern, pp. 29-46, 1994.

[Raudys 91] S.J. Raudys, A.K. Jain, *Small sample size effects in statistical pattern recognition: Recommendations for practitioners*, IEEE Trans. PAMI-13, 3, pp. 252-263, March 1991.

[Rice 92] S.V. Rice, J. Kanai, T.A. Nartker, *A difference algorithm for OCR-generated text,* in Advances in Structural and Syntactic Pattern Recognition, (H. Bunke, editor), World Scientific, 1992.

[Rice 93] S.V. Rice, *The OCR Experimental Environment, Version 3, University of Nevada*, Las Vegas, TR ISRI TR-93-04, April 1993.

[Sandberg 93] J. Sandberg, *Counting OCR errors in typeset text*, Tech. Report MITL-TR-85-93, Matsushita Information Technology Laboratory, Princeton, 1993.

[Sarkar 94] P. Sarkar, *Random phase spatial sampling effects in digitized patterns*, MS dissertation, Rensselaer Polytechnic Institute, December 1994.

[Schantz 82] H.R. Schantz, *The History of OCR*, Recognition Technologies Users Association, 1982.

[Schantz 94] H. F. Schantz, *The latest OCR developments and trends in data capture*, Today 17, 2 (Association for Work Process Improvement), pp. 34-37, November 1994.

[Shankoff 83] D. Sankoff and J.B. Kruskal, *Time warps, string edits, and macromolecules:, The theory and practice of sequence comparison*, Addison-Wesley, Reading, MA 1983.

[Smith 93] R. Smith, C. Newton, P. Cheatle, *Adaptive thresholding for OCR: A significant test*, Technical Report HPL-93-22, HP Laboratories, Bristol, UK March 1993

[Spitz 94].A.L. Spitz and M. Ozaki, *Palace: A multilingual document recognition system*, Procs. DAS94, Kaiserslautern, pp. 59-76, 1994.

[Srihari 85] *Computer Text Recognition and Error Correction*, IEEE Computer Society, 1985.

[Srihari 92] S.N. Srihari, *High-performance reading machines, Proceedings of the IEEE 80*, 7, pp. 1120-1132, July 1992.

[Stabler 94] H.R. Stabler, *Experiences with high-volume, high-accuracy document capture*, Procs. DAS94, Kaiserslautern, pp. 47-58, 1994.

[Taghva 94] K. Taghva, J. Borsack, A. Condit, S. Erva, *The effects of noisy data on text retrieval*, J. of the Am. Soc. for Information Science 45, 1, pp. 50-58, 1994.

[Trier 95] O.D. Trier, A.K. Jain, *Goal-directed evaluation of binarization methods*, IEEE-Trans. PAMI 17, to appear.

[Viswanthan 89] M. Viswanathan and M.S. Krishnamoorthy, *A Syntactic Approach to Document Segmentation*, in Structured Pattern Analysis, (R. Mohr, T. Pavlidis, A. Sanfelieu, eds.), World Scientific Publ. Co., pp. 197-215, 1989.

[Wagner 74] R.A. Wagner and M.J. Fischer, *The string-to-string correction problem*, J. ACM 21, 1, pp. 168-173, 1974.

[Watanabe 91] T. Watanbe, H. Naruse, Q. Luo, N Sugie, *Structure analysis of table-form documents on the basis of recognition of vertical and horizontal line segments*, Procs. ICDAR 91, St. Malo, pp. 638-646, 1991.

[Wilkinson 92] R.A. Wilkinson et al., *The first census optical character recognition systems conference*, NISTIR 4912, NIST, Advanced Systems Division Image Recognition Group, Gaithersburg, MD August 1992.

[Yu] Y. Yu, *A system for engineering drawing understanding*, PhD Dissertation, University of Nebraska - Lincoln, 1994.