Procs. Document Recognition and Retrieval VII,
SPIE Vol. 3967, 58-69 (invited)

Optical Character Recognition: An illustrated guide to the frontier

George Nagy[*a], Thomas A. Nartker[b], Stephen V. Rice[c]

[a]Dept. of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180
[b]Dept. of Computer Science, University of Nevada, Las Vegas, NV 89154
[c]Comparisonics Corporation, Grass Valley, CA 95945

ABSTRACT

We offer a perspective on the performance of current OCR systems by illustrating and explaining actual OCR errors made by three commercial devices. After discussing briefly the character recognition abilities of humans and computers, we present illustrated examples of recognition errors. The top level of our taxonomy of the causes of errors consists of Imaging Defects, Similar Symbols, Punctuation, and Typography. The analysis of a series of "snippets" from this perspective provides insight into the strengths and weaknesses of current systems, and perhaps a road map to future progress. The examples were drawn from the large-scale tests conducted by the authors at the Information Science Research Institute of the University of Nevada, Las Vegas. By way of conclusion, we point to possible approaches for improving the accuracy of today's systems. The talk is based on our eponymous monograph, recently published in The Kluwer International Series in Engineering and Computer Science, Kluwer Academic Publishers, 1999.

Keywords: OCR, character recognition, classification, recognition errors

INTRODUCTION

Progress in OCR has been steady if not spectacular since its commercial introduction at the Reader's Digest in the mid-fifties. After specially-designed typefaces, such as OCR-A, OCR-B, and Farrington 14B came support for elite and pica (fixed-pitch) typescripts, then "omnifont" typeset text. In the last decade the acceptance rates of form readers on hand-printed digits and constrained alphanumeric fields has risen significantly (form readers usually run at a high reject/error ratio). Many researchers now view off-line and on-line cursive writing as the next challenge or turn to multi-lingual recognition in a variety of scripts. Character classification is also a favorite testing ground for new ideas in pattern recognition, but since most of the resulting experiments are conducted on isolated characters, the results are not necessarily immediately relevant to OCR. Perhaps more striking than the improvement of the scope and accuracy in classification methods has been the decrease in cost. The early OCR devices all required expensive scanners and special-purpose electronic or optical hardware: the IBM 1975 Optical Page Reader for reading typed earnings reports at the Social Security Administration cost over three million dollars (it displaced several dozen keypunch operators). The almost simultaneous advent about 1980 of microprocessors for personal computers and of charge-coupled array scanners resulted in a huge cost decrease that paralleled that of general-purpose computers. Today, shrink-wrapped OCR software is often an add-on to desktop scanners that cost about the same as a printer or facsimile machine.

Although OCR is widely used, its accuracy today is still far from that of a second-grade child, let alone a moderately skilled typist. Even 99% accuracy translates to 30 errors on a typical printed page of 3000 characters (and it takes an alert proofreader to catch them all). In almost every application, either the OCR results must be corrected by a human operator, or a significant fraction of the documents are rejected in favor of operator entry. Our purpose in this presentation is to examine in some detail examples of the errors committed by current OCR systems and to speculate about their cause and possible remedy.

Major applications include casual home or office use (where the OCR output is typically corrected using a word processor), forms processing (bank checks and medical claims forms), address reading (mail, express, document routing), and the conversion of large archives of text (patents, legal tomes, historical documents) to computer-readable form. It was such an

archival conversion task that led to the project described here. The archive in question is the Licensing Supporting System (LSS) operated by the US Department of Energy, which contains all the documents relevant to the disposition of spent nuclear fuel. Some of these papers may be subpoenaed in the year 3000! The legal requirement to retain the information - estimated at forty million pages - spawned the Information Science Research Institute (ISRI) at the University of Nevada, which was charged to explore the feasibility of using commercial OCR systems to convert, index, and access this database.

In response to this mission, ISRI had to develop new methodology for conducting large-scale tests of commercial systems. A large sample of documents was digitized (scanned) at several spatial sampling rates and both gray-scale and bilevel quantization. The page images were manually *zoned* to demarcate columns of text. The next step was the creation of "ground-truth," an accurate transcription of each test document, against which the OCR outputs could be compared. A systematic protocol was established to guide more than forty key-entry operators to cope with non-ASCII symbols, foreign words, mathematical notation, strange layouts, and illustrations. Each document was keyed in twice, and the results compared and reconciled. Next, a master program running on a Unix platform was crafted to control a dozen PC's which ran the OCR software of the participating vendors. Perhaps the greatest technical challenge was the development of the algorithms required to compare the stream of characters produced by each OCR program with the ground truth files. Because columns, lines and words of text are not segmented reliably by OCR systems, line and character counts could not be depended upon for error detection. Nor could the position of the characters on the page, which is subject to distortion during the digitization process and not provided by many of the commercial systems. The solution finally selected was based on string matching techniques that determined the minimum number and the location of additions, deletions and substitutions necessary to correct the OCR output to match the ground truth.

From 1992 to 1996, ISRI ran a series of tests on leading OCR systems. Vendors participating in this competition had to meet the stringent criterion that their software should be operable wholly under computer control, without any manual intervention, on thousands of page images. Although some sample page images were distributed before each test, the test pages themselves were not made available. After the tests were completed (which took several weeks each year), a number of different error metrics, including character accuracy, word accuracy, numeric vs. alphabetic accuracy, and even zoning accuracy, were computed from the string-matching results. Year by year additional metrics were introduced and the tests expanded from the original Department of Energy data to English and Spanish newspaper articles, magazine articles, and business letters. In each document class, the pages were divided for reporting purposes into five "quality groups" according to the average number of errors made on them by the OCR systems as a group. (The prediction of OCR accuracy from apparent measures of page quality, such as contrast, stroke thickness, layout, typeface and type size, remains an open problem.) The experimental results were subjected to statistical analysis to determine the reliability of the error estimates and establish bounds of significance. The compiled results were presented at the Symposia on Document Analysis and Information Retrieval in Las Vegas and reported in the ISRI Annual Reports.

The symposium presentations of the test results also included pictorial examples (i.e., bitmaps) of phrases that gave trouble to the majority of the OCR systems. These "snippets" proved so popular with both industry representatives and academic researchers that we decided to make them publicly available in book form and to present some of them here. We started with a collection of about 1000 snippets (scanned bilevel at 300 dpi) on which more than one system made an error, and analyzed them to determine the likely source of the error. The three OCR systems we considered are very close to each other in terms of accuracy, both on the selected samples and on the much larger page-image database. We eventually culled our collection to 280 snippets and organized them into four major classes. Within each major class, we divided the errors into subclasses that we could identify with reasonable confidence:

Imaging Defects	**Similar Symbols**	**Punctuation**	**Typography**
Heavy Print	Similar Vertical Symbols	Commas and Periods	Italics and Spacing
Light Print	Other Similar Symbols	Quotation Marks	Underlining
Heavy and Light Print		Special Symbols	Shaded Backgrounds
Stray Marks			Reverse Video
Curved Baselines			Unusual Typefaces
			Very Large Print
			Very Small Print

Some snippets exhibit multiple likely causes of error, therefore other classifications are possible. The snippets are displayed at a 4X magnification where the individual pixels are visible. We also show the outputs of the three OCR systems, and our interpretation of the specific cause of error. In these conference proceedings, we can show only a few illustrations of each class: the whole annotated collection fills a 200-page book. The page images from which the snippets originate (in TIFF Group IV encoding), the ground truth, a set of software tools for measuring accuracy with some sample output, and a manual, are available to academic researchers for a nominal preparation and mailing cost as the

OCR-FRONTIERS TOOLKIT CD-ROM from

Dr. T. A. Nartker, Director, UNLV/ISRI - Campus Box 4021, Las Vegas, NV 89154-4021 (isri-info@isri.unlv.edu).

IMAGING DEFECTS

Imaging defects are introduced along the way between the printing process and the page image submitted for OCR. Defects may arise as soon as the slug or print head hits the paper. Porous paper causes the ink to spread, or bleed through from the verso. Coated, glossy paper does not absorb ink or toner and is liable to smudge. Very high speed printers, like newspaper presses, typically produce fuzzier type. New, heavily-inked typewriter or dot-matrix printer ribbons give rise to blotchy characters (Figure 1), while worn ribbons and printer cartridges result in faint impressions. Copying the page, especially on older copiers, results in further loss of definition. Copying the copies rapidly escalates the deterioration: even with modern copier technology, tenth-generation copies are barely legible. In our snippets, most of the imaging defects were already present in the hardcopy.

```
RecDmENSATIONS
RECO-HENATIDNS
REC-ENDATIOffS
```

Fig. 1a. The Siamese **M**'s baffle all three systems.
Small pieces are missing from the second **M**, the first **N**, and the **D**.

```
site.
sites
sitea
```

Fig. 1b. A few pixels can make all the difference. Each system recognizes the first **s**,
but only one can identify the second. Perhaps comparing the two **s**'s would help.

Nevertheless, the scanning process introduces imperfections of its own, especially in separating the print from the background. Paper is not a very high-contrast medium. The amount of light reflected from white bond paper is only about twenty times as much as that from solid, dark type. With high-contrast film, ratios of several hundred to one are achievable: that is why film is used as the master image in high-quality typesetters. Scanners are much more vulnerable than human readers to low contrast and to variations in the foreground and background reflectance of the page. Many OCR systems can adjust the scanner threshold on the basis of a preliminary scan of the page, and some can even set different values for parts of the page. But the binarization threshold was held fixed in all of our examples to reflect customary practice in data-conversion operations. On high-contrast pages (and all printing and copying processes intended for text are designed to produce high-contrast), the choice of threshold is not critical.

Gray-level scans of the image allow more detailed analysis of digital patterns and reduce the error rate on low-contrast material. It can also be argued that gray-scale increases the effective spatial resolution of the scanner, and therefore yields lower error rates on high-contrast, but small point-size or closely spaced text. As colored copy becomes more popular in the office world because of the availability of color printers and copiers, we can expect color scanners to follow suit.
The page is usually sampled both horizontally and vertically at the rate of 240 or 300 dots per inch (dpi). The trend is towards higher sampling rates, and some OCR packages can take advantage of 400 dpi or 600 dpi images. Others merely subsample or interpolate the image to 300 dpi. Hairline strokes and small openings are much less likely to be detected in text set in a small point size (6-pt or 8-pt) than in "normal" (10-12 pt) sizes. (If the threshold is set low enough to detect hairline strokes, then small white openings will be filled.) Most OCR systems also accept facsimile images in coarse or fine mode.

SIMILAR SYMBOLS

All OCR devices recognize characters primarily by their shape. Shape is, however, an elusive concept. We first take the large view and examine the general nature of shape, then focus in on the confusions between printed characters with similar shapes. We begin with typical definitions:

> SHAPE: *The outline or characteristic surface configuration of a thing: a contour; form.*
> (American Heritage Dictionary)

> SHAPE: *External form or contour; that quality of a material object (or geometric figure) which depends on constant relations of position and proportionate distance among all the points composing its outline or its external surface; a particular variety of this quality.* (Oxford English Dictionary)

In English, "shape" and "form" are synonyms. For the purpose of pattern recognition, the above definitions are of little help. We need to adopt a more operational definition:

> SHAPE is a property of both a set of objects and a particular method of observation or measurement. The universe of objects may be finite (e.g., a set of printed characters on a page), or infinite (the set of all triangles). The measurements are restricted to the geometry of the envelope (external and internal boundaries) of the objects, and only measurements that are invariant to translation and scale are admissible. Shape cannot depend on size, color, or location. Then *all objects that cannot be distinguished by the given method of observation are said to have the same shape*.

Thus we consider shape in terms of equivalence classes induced on a particular context or application (the universe of objects) by a particular system of features or measurements. Our definition does not reflect the idea that shape is a global property that is not affected by *insignificant* variations of the boundary. Indeed, the notion of significant variation is recognized to be thoroughly application-dependent. For instance, the short glyph that distinguishes the silhouette of a Q from an O has no effect when appended to a C. As stated by Richard Duda and Peter Hart in their classic text on pattern recognition: *The challenge is to find descriptions that are invariant to transformations leaving figures unaltered in "unimportant" ways, yet sensitive to transformations that change figures in "important" ways.*

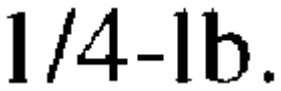
Our definition of shape rests on two pillars, the universe of objects and the method of observation. These are discussed bearing in mind an observation of Benoit Mandelbrot, the father of fractals: *...the notion that a numerical result should depend on the relation of the object to the observer is in the spirit of physics in this century and is even an exemplary illustration of it.*

Abstractly speaking, an alphabet consists simply of a fixed number of differentiable symbols. In writing and printing, each member of the alphabet is rendered by one of an infinite set of images that share only the elusive quality of shape. In the Roman, Greek and Cyrillic alphabets there are only a few letters that are represented by markedly different shapes: a familiar example is A, a, ɑ and *ɑ*. In Arabic and, to a lesser extent, in scripts derived from Sanskrit, the shape of a letter may depend on its neighbors according to a well-formed graphic grammar. (This also holds for ligatures in some Latin typefaces.)

We cannot state that a shape is rounded unless we specify how we measure curvature. If we say that an object is elongated, we must say how much it is longer than wide. If we want to describe a right angle or ell-shape (an ell is the right-angled wing of a house), we must give the minimum length of the legs. Such quantitative definitions of shape features underlie many of the methods used in OCR. In the case of printed Latin letters and numbers, shape relates to the length, width, curvature, orientation, and relative position of *strokes.* While strokes are fairly evident in handwriting, some imagination is required to decompose a typeset character into its constituent strokes.

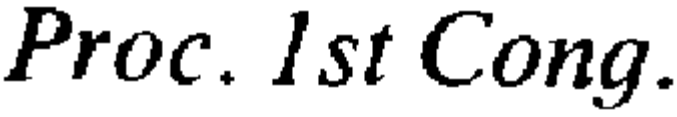
The study of invariant features that describe printed and hand-printed characters remains a topic of continuing interest. However, almost any attempt at verbal or formal mathematical definition of the shape of any particular letter will occasionally fail: there will be images that obey the description that will be instantly recognizable as belonging to another class, and images that do not obey the description that would be correctly classified by any human. An instructive illustration showing how any of the ten numerals can be deformed into any of the others through a series of continuous transformations is found in G.G.N. Wright's 1952 *The Writing of Arabic Numerals* (University of London Press).

A more subtle aspect of shape is the graphic unity of all of the alphabetic characters of a given typeface. Type designers strive to achieve such unity yet give their typeface a distinctive personality suitable for a particular context. Among the shape features used to distinguish typefaces are the aspect-ratios of the characters, the lengths of the ascenders and descenders, the ratios of the widths of horizontal, vertical and slanted strokes (which, even in a given typeface, must be altered slightly from point-size to point-size to preserve the illusion of sameness), the size and form of serifs, the eccentricity of ovals, the flare of hooks, and so forth.



I /4 lb.
l/4-lb.
1/4-lb.

Fig. lb.
Fig. lb.
Fig. lb.



Proc. I st Cong.
Proc. I st Cong.
Proc. 1st Cong.

1-57: Left lane
I-57: Left lane
1-57: Left Lane

Fig. 2. In the top example, the vertical symbol next to the **b** is probably a letter. **lb.** is probably in the lexicon, but not **Ib.** However, this logic gives the wrong result for all three systems in the figure caption! From these examples, it is difficult to tell how much each system depends on the rule that a character near a numeral is more likely to be a numeral than a letter.

The confusion between some pairs of letters, and especially between letters and numerals (Figure 2), is due to the overload imposed on written communication by cultural and scientific advances. For instance, u and v, and i and j, did not become separate letters in English until the late Middle Ages. No words of English are distinguished solely by any of these pairs, but an i-j confusion can be fatal in interpreting a computer program. Until the last century, capital letters were seldom used except for starting a sentence, and were sufficiently complex and ornate to be easily distinguished from the lower-case letters. There were few acronyms, and people were known by their whole name instead of their initials. Arabic numerals rarely appeared in text, so l-1, I-1, and O-0 confusions did not arise. Many of the special symbols such as % and $ are relatively modern inventions.

The confusion between similar shapes in contemporary printed text is easily resolved by human readers, because they don't consider each letter or numeral in isolation. In addition to drawing on context, experienced readers adapt instantly to each typeface. Type designers carefully preserving the distinction between different symbols in the same typeface. It is therefore far more likely for a symbol to resemble a different symbol in another typeface than one in its own typeface. We expect that more and more of these clues, which go well beyond the shape of individual patterns, will be exploited by OCR systems.

PUNCTUATION

Capitalization and punctuation are guideposts in written material much like inflection and phrasing in speech. In narrative and descriptive text about 60% of all punctuation consists of periods and commas, with commas more abundant than periods. However, in technical material periods outnumber commas, because in addition to bona fide punctuation they are also used in abbreviations, decimal numbers, and ellipses. The frequency of commas in written text has been dropping for centuries along with average sentence length. Three hundred years ago commas were about three times as common as periods. Considering their prevalence, it is unfortunate that commas and periods look so similar (Figure 3). Their small size prevents type designers from doing much to distinguish them.

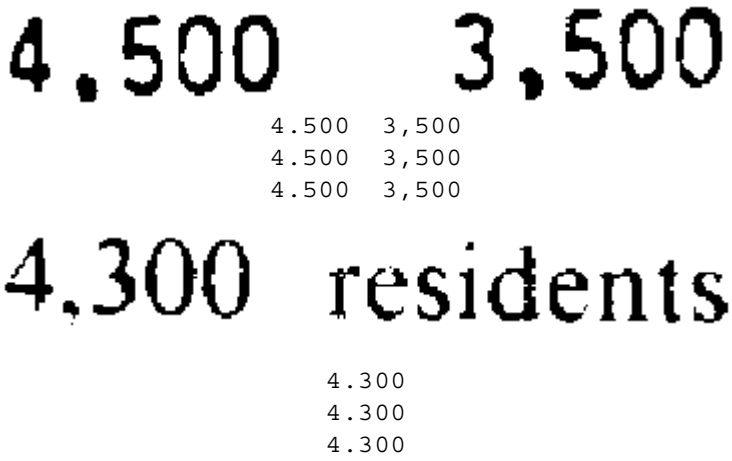


Fig 3. All three devices conspire for a thousand-fold reduction of **4,500** and **4,300**, but a few extra pixels save **3,500**.

Hyphens (-), em—dashes and en–dashes are also relatively common (15%). Some OCR systems don't distinguish between them. So are quotation marks and apostrophes (10%), parentheses (5%) and, in scientific text, brackets (used to increasingly instead of superscripts for citations) and braces. In some typefaces, apostrophes and single and double quotation marks are distinguished from commas only by their elevated position with respect to the baseline. In Spanish and French, «comillas/guillemets» are used instead of quotation marks. Exclamation marks are used only at the writer's peril! And isn't it hard to find a spot for a question mark?

Sadly, semicolons and colons account for less than 5% of the punctuation marks in contemporary writing. In the *King James Bible* and in Shakespeare's *First Folio*, more than 10% of the punctuation marks were colons. The frequencies of all types of punctuation varies considerably with the author and type of material. Many programming languages assign special meanings to punctuation symbols. The @ sign is proliferating on the Internet.

Plain English text needs only a few other symbols: slash (as in the deplorable and/or), ampersand, percent sign and, for the concupiscent, a dollar, pound or yen sign. Scholarly and technical publications, on the other hand, require an enormous number of "special" symbols. Scientists have pressed into service entire foreign alphabets, including Greek, Gothic, and Hebrew. Linguists have a plethora of phonetic symbols. Mathematicians use arcane notations for the concise expression of abstract relations, and need several levels of superscripts and subscripts to distinguish variables. Attorneys resort to a few special symbols for citations, but a greater challenge for OCR is their addiction to boldface (for insertion) and overstrike (for redaction). The European Patent Office reportedly needs over 500 special symbols for the conversion of its multilingual patent collection to computer-readable form. In the ISRI samples, punctuation and special symbols account for about 5% of the text, or half as much as the most common letter, e. Interword blanks, by comparison, are about 15%. Among the ASCII symbols, the commas, periods, quotation marks and apostrophes are most troublesome.

Typewriters used to enforce a strict limit on the number of special symbols, but with computerized document production systems, the possibilities are almost boundless: ♥♠♣◆, ✉☎✂☺. The number of combinations of black and white pixels on a 20 x 20 array is over a googol (i.e., 1 followed by 100 zeros). Unlike Chinese and Japanese OCR systems, readers designed for Western markets have a repertory of fewer than 200 symbols. In addition to the ASCII set of 96 printable characters, they may recognize the Greek alphabet, some financial symbols, and subscripts and superscripts. Most OCR vendors offer readers for the several European alphabets, all of which boast diacritical marks (accents and dots over the letters). The Scandinavian languages slash some of the letters, the French have their cedilla, and the Spaniards append a tilde on the n and turn the question mark upside down. In addition, some OCR systems have a provision to train the system on a few special characters.

TYPOGRAPHY

Here we consider the problems imposed by the *intended* appearance of printed text, as opposed to defects that arise in printing, copying, or scanning. Responsibility for the final appearance of the text is shared by the publisher, editor, graphic designer, typographer, compositor, and printer. In desktop publishing, a single person may wear all of these hats.

Typography is the art, or skill, of designing communication by means of the printed word. Good typography is transparent: it reveals rather than hides the message. The OCR user, however, has no control over the typography and is often faced with layouts, typefaces, and type sizes that are far from ideal for OCR. ("Turn-around documents," such as utility invoices bearing the name and account number of the customer, are an exception. Here the form may be expressly designed for ease of recognizing the preprinted information by OCR.)

Many of the typefaces used today are derived from medieval calligraphy, only slightly modified by the limitations of early printing technologies (wood blocks and moveable type). To imitate fine penmanship, vertical strokes are thick compared to horizontal strokes, and NW-SE diagonals are thicker than NE-SW diagonals. Therefore curved strokes may vary in width according to their local orientation. In contrast, typefaces specifically designed for accurate OCR, such as OCR-A and OCR-B fonts, have uniform stroke widths and exaggerated distinctions between similar symbols such as O and 0 or 1 and l.

At first, only cast metal type was used. For centuries after Gutenberg, the type was cut, molded, cast and set by hand. After the First World War this laborious process was superseded by keyboard-driven composition and typesetting such as the *Monotype* and *Linotype* systems. Lithography was invented as early as 1798, but until recently was used only for fine books. As the name suggests, lithography originally exploited the affinity of certain types of stones for ink, but now denotes any kind of plate printing without raised type. Lithography allows longer press runs and greater ease in reproducing illustrations than hot type. Ink-on-stone eventually led to offset processes using metal or plastic plates for high-quality color printing (*chromolithography*). While offset-lithography still requires metal type to produce printing plates, *filmsetting* is accomplished by the computer-controlled light beam of a photocomposition machine. This technique, only a few decades old, is now being replaced by 'direct printing' xerographic techniques. Laser-printers for desktop publishing have already shifted from 300 dots-per-inch (dpi) to 600 dpi, while commercial photocomposers and xerographic printers operate at 1200 to 2400 dpi.

Moveable type had the greatest impact on alphabetic (as opposed to ideographic) writing systems such as Latin and Greek. The English alphabet has only 26 *letters*, but about 80 *characters* are commonly used in non-technical material (upper and lower case, numerals, punctuation, and some special symbols). These are all included in the ASCII (American Standard Code for Information Interchange) set of 96 characters adopted by US computer manufacturers. The Unicode offers hope that

eventually all the symbols in the writing systems of the whole world - including Chinese and Japanese - will be represented without ambiguity. Although even in English several hundred symbols may be required for specialized publications like patents, dictionaries, and scientific texts, we tested OCR systems using only the ASCII set.

Aesthetic considerations dictate a certain consistency in the shape of the letters and numerals. A group of related shapes that represent all the customary symbols is called a *typeface*. Popular examples are Times New Roman, Bookman, Bodoni, Futura. Not all typefaces with the same name are identical: for complete identification of a typeface, one must add the name of the supplier, e.g., Bauer Futura. Each typeface offers several stylistic variants, e.g. *italics* and SMALLCAPS, and weights (light, medium, **bold, extrabold**), whose use is governed by long-established conventions. In addition, there may be *condensed* and *expanded* versions that give additional flexibility for page layout. In conventional typography, a *font* (derived from 'found', as in type foundry) was a complete set of letters and other symbols, each in approximate proportion of frequency of usage, in a single body-size and design. Now it is commonly used, especially in word-processing circles, to describe a single typeface-size-style combination. Type size is measured in *printer's points*: one point is 1/72 of an inch. For text set in paragraphs (*body text*), 9 to 12 point type is normally chosen for greatest legibility, and these sizes account for well over 95% of all the samples in the ISRI tests.

A Typ I (Association Typographique Internationale) has adopted an elaborate taxonomy (the Vox system) for classifying all typefaces into twelve major categories. For our purposes, it is sufficient to differentiate between *serif* and *sans-serif* (sometimes called *lineal*) typefaces and between *variable-pitch* and *fixed-pitch* type. We call unusual typefaces, designed more for attracting attention than for legibility, *display type*. This category includes the shaded, outline, and reverse-video characters that are often used in advertising, letterheads, and headlines.

The word 'roman' can mean a type with serifs, or the opposite of 'italic' or, for numerals, figures set in capital letters rather than Arabic numerals. 'Italic' also has several meanings. Some authors reserve 'italic' for serif types, and use 'oblique' for slanted sans-serif. Others consider 'italic' a synonym for cursive fonts designed to imitate handwriting. Script typefaces are slanted but are not intended as a stylistic variant for a corresponding upright body type.

Smaller sizes are used for footnotes, and larger sizes for titles and headings. Note that the size in points indicates only the size of the metal body: the height of the characters appearing on paper may vary according to the typeface. More useful in OCR are the relative distances between the *ascender line*, the *x-line*, the *base-line*, and *the descender line*. The height of an upper-case ten-point letter scanned at 300 dpi is about 24 pixels, while the dot on the i may have a diameter of only 4 pixels.

The horizontal size of type is called the *set*, and is also measured in points. The width of the characters in a normal typeface varies by a ratio of over 3:1 from m to i. The variation in type width increases legibility because it differentiates word shapes, but numerals are often of uniform width to facilitate tabular composition. In professional typesetting, words are separated only by the width of an i. *Justification* (even right margins) is achieved by changing the spacing between letters of the same word as well as between words, and occasionally by judicious (and almost imperceptible) use of condensed or expanded type.

Kerning is a means of achieving a pleasingly uniform darkness in a line of text (Figure 4). To avoid offensive white spaces, an x-height character such as e or o may be slipped under the wing of a V or W, or a period nestled under the protruding chest of a P. A pair of letters is kerned if they cannot be cleanly separated by a vertical line. In high-quality printing, italics and script are always kerned.

Some frequently occurring pairs of letters, such as fi, fl, ffi, ffl are designed in many typefaces as a single shape, called a *ligature*. Ligatures may be considered an extreme case of kerning. Because a ligature cannot easily be decomposed into recognizable constituents -- for instance, the dot of the i in fi is absorbed by the f -- OCR systems just consider them as a separate symbol. Upon identifying a ligature, the OCR system puts out the appropriate ASCII character codes.

Entire words may be created as a single highly-stylized graphic unit, called a *logotype*. Logotypes may be registered as trademarks. In OCR applications they appear most often on letterheads. Automating the recognition of logotypes is an on-going area of research.

Fig. 4. Kerned italics cause a variety of problems, include the omission of interword blanks. Underscoring is also difficult. All the devices are bewildered by the enormous flourish on the capital **R** (truly a 'majuscule'), and by the z-like appearance of the lower-case **r**.

The spacing between lines of text (*leading*, or *inter-linear spacing*) is also expressed in points. In solid-set paragraphs, there is no spacing whatever between blocks of type. Type set 10/14 means that there is a 4-point vertical space between the blocks of 10-pt type. Good legibility, and also accurate OCR, requires at least 2-pt spacing. Underscores are, of course, seldom used in printed text, but are an accepted typewritten substitute for italics (Figure 4).

In summary, the graphic coding system we use to represent language in visual form is complex but well-suited to human abilities. Because it is woven so deeply into the fabric of our culture, it is unlikely to undergo rapid change. We shall not learn to read bar codes. We must therefore depend on OCR to eradicate computer illiteracy. In the meantime, unusual typefaces still confound OCR systems. Even the largest OCR training sets exhibit very limited variety compared to the ordinary reader's repertory.

CONCLUSION

We have seen that, for a variety of reasons, current OCR devices cannot even read as well as a seven-year old child. In order to extend the scope of prospective applications to more difficult material, researchers in many countries are hard at work to develop better methods for reading printed text. We speculate on what might be in the wings, but try to keep an open mind about the nature of the solutions that might emerge. We consider four potential sources of improvement:

- Improved image processing;
- Adaptation of the classifier to the current document;
- Multi-character recognition;
- Increased use of context.

The first method is based on more faithful modeling of the printing, copying and scanning processes. The second specializes character classification to the typeface of the current document, while the third exploits style consistency in typeset text. The fourth method depends only on linguistic properties, and varies from language to language. While at first blush these four notions seem very different, we shall see that the borders are fuzzy.

Image processing

Degraded copy can often be traced to facsimile, old copiers and computer printers, poorly maintained typewriters, high-speed tabloid presses, coarse paper, or heavy handling. Improved image processing aims at alleviating the effects of printing defects like stray marks and curved baselines, and of certain typographic artifacts like underlining, shaded background and reverse video. Some of the necessary techniques have already been developed, but are still too resource-intensive for the current generation of desktop systems. Even simple filtering algorithms require processing a 3x3 or 5x5 window centered on every single pixel in the image. The detection of large interfering marks - blots, underscores, creases - requires even more processing power. Equally time-consuming is the localized skew correction that is necessary for handset pages and for baseline curl on pages copied from bound volumes.

Faster processors will also allow gray-level and color scanning at spatial sampling rates higher than 300 dpi. Although this may be difficult to judge from our snippets, there is evidence that the processing of low-contrast documents with small print can be improved by multi-level gray-scale quantization and a higher spatial sampling rate.

Gray-tone scanning can be used either for adaptive binarization or for gray-scale feature extraction. Adaptive local binarization helps cope with uneven contrast, but fine or faint connecting strokes can be more easily detected by complete gray-scale processing. High-resolution spatial sampling reduces the edge effects due to the unpredictable location of the sampling grid, which is often the dominant source of noise in very small print. Color scanning will eliminate problems due to colored print and shaded backgrounds. Sound but resource-intensive techniques are already available for dealing with halftones.

Our most flagrant examples of light and heavy print are caused by high-contrast, low-resolution copying rather than by poor thresholding during scanning. Gray-level scanning will not help here. The thinning and thickening of strokes is caused by the combined effect of the unknown point spread functions and non-linear intensity quantization in the printing, copying and scanning processes. Determining this effect and compensating for it for every document appears to be a very difficult task.

There have been some attempts at detailed characterization of the types of noise found in printed images, in the hope that such noise models will allow the generation of immense character-image data sets for training noise-immune classifiers. It is possible that further development of pseudo-defect character and page-image models will lead to classifiers that are less sensitive to point-spread distortion and other types of imaging noise. We place, however, greater hope in the preservation of gray tones and fine detail.

Adaptation

Adaptation means exploiting the essentially single-font nature of most documents. (Typeface catalogs and ransom notes pasted together with letters from many sources are relatively rare). Italics and boldface can be considered simply as part of an extended typeface. Single-font classifiers designed for a particular typeface are far more accurate (and faster) than multifont or omnifont classifiers. There is good reason for this: type designers are careful to maintain distinctions between glyphs that represent the different symbols within the same typeface. However, a glyph in one typeface (for instance an "ell", might be identical or very similar to a glyph representing a different symbol (an "eye" or a "one") in another typeface. Errors of this type are not confined to tall slender characters: one man's fat "Oh" is another man's rotund "Zero."

Such inter-font confusions can be avoided by restricting the classifier to the appropriate choice in the current font only, either by adapting the classifier parameters to the document, or by automatic font identification. The former requires that at least some samples of each character be correctly identified by the initial (multifont) classifier. The latter can be based on features that are common to many characters in the same typeface: size and shape of serifs, the ratio of ascender and descender height to x-height, the angle of slanting strokes, and the ratio of the widths of horizontal and vertical strokes. Unlike classifier adaptation, font identification requires a vast storehouse of every possible font under every possible imaging condition. Consequently we believe that classifier adaptation is the more promising approach.

A low-risk approach to adaptation can make use of the normal quality control process. In most OCR installations, there is a great deal of similarity among the documents processed from day to day. Since the mistakes made by the OCR system are often corrected by an operator, a significant amount of training data becomes available after every shift. This data is far more likely to be representative of future loads than any factory design data. The data can therefore be used for automated retraining of the OCR system during low-activity periods. Even if the classifier does not improve from day to day, at least it will not persist in making the same mistake.

Multi-character recognition

Instead of recognizing individual characters, it may be desirable to recognize the bit-mapped images of larger units. Among the resulting benefits is avoidance of error-prone character-level segmentation. One option is to recognize complete words, under the assumption that inter-word blanks can be readily detected. Entire words typically have a characteristic contour based on the relative location within the word of letters with ascenders and descenders. For instance, `pebbles` should not be confused with `Cables`. Among several lexicon words with the same outline, recognizing just one or two of the letters usually suffices to disambiguate them.

A few hundred common words - like *the*, *a*, *an*, *to*, *from* - account for over one half of all the words in normal English text. Once these common words (called "stop words" in information retrieval) have been recognized, they can be analyzed to discover the exact shape of their constituent letters. This helps, in turn, to recognize the remaining words. Thus word recognition can lead to adaptation.

By programming the recognition of more informative words instead of stop words, whole-word recognition can be applied to keywords in information retrieval tasks. However, techniques based on word outlines fail on all-cap text, on text that is not composed of meaningful words (like industrial part numbers) and on numeric fields. Furthermore, the number of common word forms is much higher in languages like Italian with extensive declensions, conjugations, and gender agreements.

The units to be recognized need not be entire words: they could simply be pairs of glyphs. The advantage is that while `I` ("I") cannot be recognized unambiguously because `I` ("l") may have the same shape in another font, the combination `It` ("It") may differ from `lt` (`"lt"`). Multi-character recognition can be accomplished systematically by processing the pixels that appear in a moving window of several character-widths. The appropriate mathematical foundations - Hidden Markov Models - have already been worked out for cursive writing and for speech recognition, and the resulting algorithms are now beginning to be applied to printed text. The most ambitious of these algorithms can recognize an entire page image in one fell swoop.

None of these methods work when identical glyphs that represent different symbols within the same font. This happens on some old typewriters, where 1 and l, and 0 and O, were on the same key. To resolve such ambiguities, we must resort to context.

Linguistic context

In current OCR systems, the use of context is restricted to choosing a common letter n-gram (like `ing`) over a rare one (`lng`), or a word that appears in the lexicon over one that does not (`"bolt"` rather than `"holt"`). Just making use of word frequencies would improve recognition. Customizing the lexicons to a particular application or a set of documents could cure more of the problems that are caused by an inappropriate vocabulary. As pointed out by Professor Pavlidis, in a business letter, `"date"` is more likely than `"dale"`, while the opposite is true in a pastoral elegy. Future OCR systems will have a better sense of the right word.

Each letter is constrained not only by its neighbors within the same word, but also by the neighboring words. This takes us beyond morphology and lexical analysis to syntax and semantics. While most of our snippets are too short to demonstrate the potential power of syntactic analysis, we do have some examples. For instance, `"leek tar fail, tar lent"`, while lexically acceptable, is clearly ungrammatical. Simple grammatical rules, which have not yet been fully exploited in OCR, also govern capitalization, the intermingling of letters and numerals, and punctuation (i.e., balancing parentheses or quotation marks). In recent years, stochastic grammars based on the statistical analysis of word transition frequencies in large corpora have gained favor over traditional rule-based grammars. Stochastic grammars can be automatically compiled for any language or application for which there exists a significant body of literature in computer readable form.

Semantic analysis would allow the OCR system to avoid meaningless yet grammatically correct interpretations like `"wages of lake level."` Of course, the effectiveness of all linguistic context is limited by the possibility that the text contains unexpected constructs, including nonsense words, ungrammatical phrases, and sentences that defy semantic interpretation. Non-technical material, like a novel, typically has fewer such irregular constructs than technical material (especially books about OCR). Although human operators seldom have to resort to complex analysis to transcribe printed text, when presented with unfamiliar handwriting they unquestionably (if subconsciously) take advantage of high-level linguistic context.

Before we leave context, it is appropriate to emphasize the difference between techniques based on font, glyph-pair and word-shape on one hand, and the use of letter n-grams, lexicons and syntax on the other. Both are examples of context, but the first governs the relationship of glyphs or graphical shapes, while the second constrains the sequences of symbols regardless of the shape of the corresponding glyphs. The first set of relations is sometimes called graphical context or "style," while the second is "linguistic context".

The four approaches mentioned above are by no means mutually exclusive. Not only can they be combined within the same system, but the output of complete OCR systems, each based on entirely different principles, can be combined - as has already been demonstrated - for enhanced accuracy. This requires synchronizing the output of the classifiers, and letting each vote for the identity of the current character. As we have noted, different classifiers don't necessarily make the same mistake, so this democratic process effectively lowers the error rate.

On the basis of the diversity of errors that we have encountered, we are inclined to believe that further progress in OCR is more likely to be the result of multiple combinations of techniques than on the discovery of any single new overarching principle. We hope that our collection of snippets, demarcating the boundary of what has already been achieved, will help to extend OCR territory.

REFERENCE

S.V. Rice, G. Nagy, T.A. Nartker, Optical Character Recognition: An illustrated guide to the frontier. Kluwer Academic Publishers, 1999. (this book contains an annotated bibliography and an index to the source of each sample).