G. Nagy, Foreword to Handbook of Document Image Processing and Recognition (David Doermann & Karl Tombre, editors), Sprinter, 2014.

FOREWORD (V2)

In the beginning, there was only OCR. After some false starts, OCR became a competitive commercial enterprise in the 1950's. A decade later there were more than 50 manufacturers in the US alone. With the advent of microprocessors and inexpensive optical scanners, the price of OCR dropped from tens and hundreds of thousands of dollars to that of a bottle of wine. Software displaced the racks of electronics. By 1985 anybody could program and test their ideas on a PC, and then write a paper about it (and perhaps even patent it).

We know, however, very little about current commercial methods and in-house experimental results. Competitive industries have scarce motivation to publish (and their patents may only be part of their legal arsenal). The dearth of industrial authors in our publications is painfully obvious. Herbert Schantz's book, *The History of OCR*, was an exception: he traced the growth of REI, which was one of the major success stories of the 1960's and 1970's. He also told the story, widely mirrored in sundry wikis and treatises on OCR, of the previous fifty years' attempts to mechanize reading. Among other manufacturers of the period, IBM may have stood alone in publishing detailed (though often delayed) information about its products.

Of the 4000-8000 articles published since 1900 on character recognition (my estimate), at most a few hundred really bear on OCR (construed as *machinery* - now software - *that converts visible language to a searchable digital format*). The rest treat character recognition as a prototypical classification problem. It is, of course, researchers' universal familiarity with at least some script that turned character recognition into the pre-eminent vehicle for demonstrating and illustrating new ideas in pattern recognition. Even though some of us cannot tell an azalea from a begonia, a sharp sign from a clef, a loop from a tented arch, an erythrocyte from a leukocyte, or an alluvium from an anticline, all of us know how to read.

Until about 30 years ago, OCR meant recognizing mono-spaced OCR fonts and typewritten scripts one character at a time – eventually at the rate of several *thousand* characters per second. Word recognition followed for reading difficult-to-segment typeset matter. The value of language models more elaborate than letter n-gram frequencies and lexicons without word frequencies gradually became clear. Because more than half of the world population is polyglot, OCR too became multilingual (as Henry Baird predicted that it must). This triggered a movement to post all the cultural relics of the past on the Web. Much of the material awaiting conversion, ancient and modern, stretches the limits human readability. Like humans, OCR must take full advantage of syntax, style, context, and semantics.

Although many academic researchers are aware that OCR is much more than classification, they have yet to develop a viable, broad-range, end-to-end OCR system (but they may be getting close). A complete OCR system, with language and script recognition, colored print capability, column and line layout analysis, accurate character/word, numeric, symbol and punctuation recognition, language models, document-wide consistency, tuneability and adaptability, graphics subsystems, effectively embedded interactive error correction, and multiple output formats, is far more than the sum of its parts. Furthermore, specialized systems - for postal address reading, check reading, litigation, and bureaucratic forms processing - also require high throughput and different error-reject trade-offs. Real OCR simply isn't an appropriate PhD dissertation project.

I never know whether to call hand print recognition and handwriting recognition "OCR." but abhor *intelligent* as a qualifier for the latest wrinkle. No matter: they are here to stay until tracing glyphs with a stylus goes the way of the quill. Both human and machine legibility of manuscripts depend significantly on the motivation of the writer: a hand-printed income tax return requesting a refund is likely to be more legible than one reporting an underpayment. Immediate feedback, the main advantage of on-line recognition, is a powerful form of motivation. Humans still learn better than machines.

DIA is a superset of OCR, but many of its other popular subfields require OCR. Almost all line drawings contain text. An E-sized telephone company drawing, for instance, has about 3000 words and numbers (including revision notices). Music scores contain numerals and instructions like *pianissimo*. A map without place names and elevations would have limited use. Mathematical expressions abound in digits and alphabetic fragments like *log, limit, tan* or *argmin*. Good lettering used to be a prime job qualification for the draftsmen who drew the legacy drawings that we are now converting to CAD. Unfortunately, commercial OCR systems, tuned to paragraph-length segments of text, do poorly on the alphanumeric fragments typical of such applications. When Open Source OCR matures, it will provide a fine opportunity for customization to specialized applications that have not yet attracted heavy-weight developers. In the meantime, the conversion of documents containing a mix of text and line art has given rise to distinct sub-disciplines with their own conference sessions and workshops that target graphics techniques like vectorization and complex symbol configurations.

Another subfield of DIA investigates what to do with automatically or manually transcribed books, technical journals, magazines and newspapers. Although Information Retrieval (IR) is not generally considered part of DIA or vice-versa, the overlap between them includes "logical" document segmentation, extraction of tables of content, linking figures and illustrations to textual references, and word spotting. A recurring topic is assessing the effect of OCR errors on downstream applications. One factor that keeps the two disciplines apart is that IR experiments (e.g., TREC) typically involve orders of magnitude more documents than DIA experiments because the number of characters in any collection is far smaller than the number of pixels.

Computer vision used to be easily distinguished from the image processing aspects of DIA by its emphasis on illumination and camera position. The border is blurring because even cellphone cameras now offer sufficient spatial resolution for document image capture at several hundred dpi as well as for legible text in large scene images. The correction of the contrast and geometric distortions in the resulting images goes well beyond what is required for scanned documents.

This collection suggests that we are still far from a unified theory of DIA or even OCR. The Handbook is all the more useful because we have no choice except to rely on heuristics or algorithms based on questionable assumptions. The most useful methods available to us were all *invented* rather than derived from prime principles. When the time is ripe, many alternative methods are invented to fill the same need. They all remain entrenched candidates for "best practice". This Handbook presents them fairly, but generally avoids picking winners and losers.

"Noise" appears to be the principal obstacle to better results. This is all the more irritating because many types of noise (e.g. skew, bleed-through, underscore) barely slow down human readers. We have not yet succeeded in characterizing and quantifying signal and noise to the extent that communications science has. Although OCR and DIA are prime examples of information transfer, information-theoretic concepts are seldom invoked. Are we moving in the right direction by accumulating empirical midstream comparisons – often on synthetic data – from contests organized by individual research groups in conjunction with our conferences?

Be that as it may, as one is getting increasingly forgetful it is reassuring to have most of the elusive information about one's favorite topics at arm's reach in a fat tome like this one. Much as on-line resources have improved over the past decade, I like to turn down the corner of the page and scribble a note in the margin. Younger folks, who prefer search-directed saccades to an old-fashioned linear presentation, may want the on-line version.

David Doermann and Karl Tombre were exceptionally well qualified to plan, select, solicit, and edit this compendium. Their contributions to DIA cover a broad swath and, as far as I know, they have never let the song of the sirens divert them from the muddy and winding channels of DIA. Their technical contributions are well referenced by the chapter authors and their voice is heard at the beginning of each section.

Dave was co-founding-editor of *IJDAR*, which became our flagship journal when *PAMI* veered towards computer vision and machine learning. Along with the venerable *PR* and the high-speed, high-volume *PRL*, *IJDAR* has served us well with a mixture of special issues, surveys, experimental reports, and new theories. Even earlier, with the encouragement of Azriel Rosenfeld, Dave organized and directed the Language and Media Processing Laboratory, which has become a major resource of DIA data sets, code, bibliographies, and expertise.

Karl put Nancy on the map as one of the premier global centers of DIA research and development. Beginning with a sustained drive to automate the conversion of legacy drawings to CAD formats (drawings for a bridge or a sewer line may have a lifetime of over a hundred years, and the plans for the still-flying Boeing 747 were drawn by hand), Karl brought together and expanded the horizons of University and INRIA researchers to form a critical mass of DIA.

Dave and Karl have also done more than their share to bring our research community together, find common terminology and data, create benchmarks, and advance the state of the art. These big patient men have long been a familiar sight at our conferences, always ready to resolve a conundrum, provide a missing piece of information, fill in for an absentee session chair or speaker, or introduce folks who should know each other.

The DIA community has every reason to be grateful to the editors and authors of this timely and comprehensive collection. Inexpensive illegitimate offprints should soon be available. Enjoy, and work hard to make a contribution to the next edition!

George Nagy
Professor Emeritus, RPI