Digitizing, Coding, Annotating, Disseminating, and Preserving Documents

George Nagy
ECSE DocLab, Rensselaer Polytechnic Institute, Troy, NY, USA 12180
nagy@ecse.rpi.edu

ABSTRACT

We examine some research issues in pattern recognition and image processing that have been spurred by the needs of digital libraries. Broader – and not only linguistic – context must be introduced in character recognition on low-contrast, tightly-set documents because the conversion of documents to coded (searchable) form is lagging far behind conversion to image formats. At the same time, the prevalence of imaged documents over coded documents gives rise to interesting research problems in interactive annotation of document images. At the level of circulation, reformatting document images to accommodate diverse user needs remains a challenge.

Categories and Subject Descriptors

I.7.5 [**Document and Text Processing**]: Document Capture

General Terms

Document processing, Preservation

Keywords

Document coding, Document administration, Digitization, Archiving

1. INTRODUCTION

The conversion of the world's accumulated hardcopy to computer-accessible formats is only a transient problem, but it is a *huge* transient problem. We consider research opportunities to help solve it, focusing on the five issues mentioned in the title. There are two common threads shared by all of them: *selection* and *quality control*. Both are heavily influenced by socio-political and economic considerations. For example, the politically correct word for selection is *prioritization*: if funding is not available for everything, what should be processed first? Quality control bears on this because it is a matter of trade-offs: is it better to process more material, sooner, and tolerate a lower quality product? Here we consider only the underlying technical requirements, although most research issues are dwarfed by the impact of the related policy decisions.

IWRIDL-2006 Kolkata, India

2. DIGITIZATION

We reserve the term *digitization* for the conversion of information on paper, papyrus, silk or photographic film to digital images. Currently unbound pages are digitized with rollerfeed scanners, and bound volumes and delicate materials with flatbed or overhead scanners. (Overhead scanners combine electronic and mechanical scan with camera optics). The visual quality of carefully scanned pages already matches that of photographic facsimiles. We expect a transition from scanners to digital cameras because precise control of mechanical motion is expensive.

Every downstream process is affected by the quality of the digitization. Two technical issues related to quality control are (1) how to specify it, and (2) how to ensure it. Geometric specifications are easy: spatial sampling rate (ppi or lpm), local and global linearity, and maximum skew. However, photometric characteristics are much more difficult to specify and measure. They cover spectral response, point-spread function, photometric transfer function, amplitude quantization, and the spatial and temporal uniformity of all four. The requirements for readability are more permissive than for scholarly research on archival materials, because the latter often focuses on minute changes in letter formation and on subtle characteristics of the ink and substrate.

In volume digitization, each page is usually inspected by an operator, who catches the most egregious degradations. Appropriate displays of on-line analyses (like gray-value and RGB histograms) can trigger an immediate rescan. Good practice also requires periodic insertion of calibration targets into the workflow. Calibration data reveals any drift in operating conditions and facilitates subsequent correction of reversible distortions (possibly making use of algorithms that haven't even been discovered yet). Scan-quality issues were discussed by a working group at DIAL04. Their recommendations were presented at DIAL06, and are included in the Proceedings thereof.

Some current research addresses the question of what scanning parameters can be recovered from scanned text and line art, *without* special calibration targets. In the library context, current scanning technology and adequate calibration render image *enhancement* unnecessary. Another fading research objective is *compression* of scanned text. After a gradual rise through three decades, compression of scanned

text is approaching the Shannon limit of a few bits per symbol. This is not surprising, because current recent document compression methods like JBIG2 and DjVu are based on glyph recognition. (The digitized image must be retained for human readers because even advanced OCR techniques cannot encode perfectly the *appearance* of the original document; therefore documents rendered from the encoded version tend to be less readable.)

3. CODING

Coding, the conversion of raster images to symbolic form, is a richer territory for research than digitization. We consider only the coding of material meant to convey *symbolic* information, like text and line drawings, as opposed to *natural* pictures like paintings and photographs. (There are, to be sure, many interesting problems associated with the latter, but they are not primarily coding problems.) After much competition between alternative codes, we are approaching worldwide agreement, or at least interoperability, in multilingual representation. Early computer folks would surely consider the length of current codes (like Unicode) profligate, but compared to music and video, the world's symbolic holdings require only negligible amounts of storage.

The conversion of plain text is the province of character recognition, about which many, many thousands of learned papers have already been written. There is no sign of abating interest: new ideas for pattern recognition are often most easily tested on character classification. Nevertheless, optical character recognition still isn't nearly good enough for many applications. Only on high quality images of text does it rival human data entry in accuracy (although it is, of course, much faster). Most current text conversion projects still require either manual data entry, or time-consuming proofreading and correction.

We believe that the key to improved OCR is the recognition of *long fields* of characters *simultaneously*. Adaptation, style-conscious classification, and clause or sentence level language models exemplify aspects of *contextually constrained field classification* that have not yet been sufficiently studied and exploited. Furthermore, extracting information from human corrections and using it to re-estimate classifier parameters for the current flow of data will be more effective than training classifiers on huge (but seldom representative) collections of labeled characters.

Most document image analysis (DIA) requires OCR at some point, but the emphasis in DIA is often on other problems, like logical page segmentation, table interpretation, or line-drawing analysis. All of these are fertile grounds for research, because in operational conditions they are still usually performed manually. The big change from 20 years ago is that the digitized documents can now be displayed on the computer screen. Selected portions can be magnified, related data from other sources can be conveniently retrieved and consulted, and information entered by the operator can be superimposed on the document image to assist navigation.

Further automation of the entire process is possible even at the current state of the art. Existing algorithms can take over many low-level functions, so that the operator need only confirm or correct the results. The coding of engineering drawings ("vectorization") is already on this path. What has not yet been done is *closing the loop*. Every operator interaction should result in some change in the configuration of the system that decreases the likelihood that the same situation will require the same intervention again. In other words, *the system must improve with use*.

4. ANNOTATION

Annotation covers a multitude of metadata potentially associated with library items. Lowlevel annotation is usually carried out by technicians, intermediate level by librarians, and high-level by domain specialists. The simplest annotation merely assigns a serial number to a digitized or coded file to maintain its correspondence with the same number stamped on the physical source document. Page counts, front matter, and process identification (e.g. scanner settings, date, and operator) may also be keyed or downloaded into the database. More elaborate annotation includes catalog entries, which in some digital libraries are still heavily influenced by the limitations of 3x5 inch cards. Such conventional catalog entries are adequate only for items intended for leisure reading. At the high end, annotation may mean a multi-volume scholarly analysis of the source item. We can also consider *language translation* (often partially automated) as high-level annotation.

Annotation is particularly important for digitized but not coded items, because images of text cannot be easily searched automatically. Such annotation (*gloss*) may be carried out, at increasing expense, at the document, page, paragraph, and line or word level. An obsolete form of annotation is the *concordance*, which is an index of all the pages on which a given word or phrase occurs. While the preparation of concordances used to be a respected scholarly activity, on coded documents the FIND feature in any word processor now does the job. Automated *document summarization* rivals human summarization, if only because the latter is often poor. *Document categorization* (including *screening* for specific applications) is also being automated. However, many types of annotation require information that is not contained in the document itself (e.g., the author's date of birth). Although such information is often available in digital form, finding it and bringing it automatically to the annotator's attention is a challenging research task.

The entry of published technical data into integrated databases is a form of annotation quickly gaining in importance. Many current research papers are not publicly available in coded form, even though they were prepared and published with computer systems. In molecular biology and in other disciplines with sky-rocketing publication rates highly qualified curators (often with PhD degrees) scour research articles to extract factual and quantitative data for populating specialized research databases. Since the curators view the documents on a computer screen, annotation is already interactive, but there are many opportunities to improve the effectiveness and speed of the interaction. One simple example is text-image word-wrapping, which facilitates multiple window placement (as browsers do with HTML pages). Another is automated comparison of graphs or tables to discover related information. In the long run, however, researchers themselves may be required to tag their published facts for automated data mining, much as they must already attach keywords for automated information retrieval. In fields where most of the instrumentation is computerized, there may not even be any fundamental reason to interpose human roadblocks between computers that produce scien-

tific information and computers that analyze it.

Annotation for locating entire documents (like subject, author and title catalogs), and specific content with a document (like concordances), will most likely be replaced by search engines. In order to rival the human reference librarian, search engines will have to grow beyond today's word, citation and linkage based capabilities. It is believed that ontologies can lead to *understanding* of document contents, but current ontologies are still too small, subject-specific, and insular.

5. DISSEMINATION

All of the activities mentioned so far are internal to the library community. Libraries are, however, public repositories which must provide access to "end users. " Indeed, one of the major *raisons d'être* for digital libraries is that dissemination of library contents in electronic form is easier and cheaper than in their original physical form. However, electronic dissemination entails many unsolved issues of standardization and cost recovery. Furthermore, dissemination may have to be compartmentalized by the clients' age, payment record, security clearance, or interests. (Conventional libraries also had to keep track of the *order* in which high-demand items were requested.)

The biggest technical challenges lie in providing both content and metadata in a variety of useful formats. Library patrons may want to print selected library contents, peruse them on projector, desktop, laptop, pocket computer, or cell-phone displays, or perhaps listen to them while driving. Some will have good vision, others may be purblind. Commercial demand has already led to useful multi-resolution techniques for road and street maps. Experiments show that synthesized audio output for plain text is acceptable even if it is not comparable to human speech. Reformatting page layout for different display sizes and formats is an active research area in which impressive results have been achieved, but challenges remain in reformatting diagrams and tables.

Library interfaces should be simplified and made more responsive to automatically created user profiles (as are search engines). Client access to many digital libraries requires excessive technical expertise and patience. The difficulty is exacerbated by the need for interoperability with intermediate repositories, which may themselves be libraries with their own retrieval and formatting conventions. Interestingly, it is commonly reported that research papers are easier to find with a standard search engine than through digital library portals and interfaces.

We cannot take it for granted that accommodating human readers remains a priority objective. Automated readers (crawlers, search engines, and autonomous agents) are rapidly becoming the prime customers of technical collections. Digital libraries must accommodate *them* (as commercial web page designers do) with appropriate metadata, even as library budgets are becoming tighter and tighter. We can only hope some libraries will continue to serve clients with the desire, skill and patience to actually *read* books.

6. PRESERVATION

Preservation is seldom of concern with mass-produced contemporary publications because cultural importance increases with age and scarcity. Most current hardcopy products lack the durability of books and manuscripts produced centuries ago, and it is difficult to look far enough ahead to see their eventual historical value. Nevertheless, most institutions – government agencies, publishers, universities, and some public figures – conscientiously archive some of their current records.

Few librarians or archivists accept *any* digital representation as a permanent substitute for the original artifact. On the other hand, most end-users may be as satisfied with an electronic facsimile as with a high-quality (and far more expensive) physical copy. Digitization can justify restricting access to the original materials, which in turn makes it easier to store them in climate-controlled vaults.

Merely digitizing or coding something does not guarantee permanent access. For instance, many records from WWII were kept on punched cards. Not only did the punch cards deteriorate, but the card readers have disappeared. Magnetic tape and disk and optical media also have a shorter life span than paper. Furthermore, the software required to read the coded data may be incompatible with computers of another generation. It is not uncommon for engineering drawings prepared on earlier Computer Aided Design systems to be rescanned and revectorized, simply because the CAD software can no longer run on any available computer. Diskette drives, tape cassette readers, and ZIP drives are disappearing. Because of the relatively short lifetime of digital media, until recently many organizations opted for archiving documents on microfilm or microfiche instead of digital media. However, at current storage costs, it is plausible to keep everything *on line*. When the server is replaced, everything is copied, so there is no need to worry about removable media forgotten in some cabinet.

A major cost of digitization is bringing the item to be digitized to the scanning facility and returning it to its normal place (or vice versa – some expensive folios are photographed in situ). It is therefore sensible, and not inordinately costly, to scan cultural artifacts at higher spatial and grey-scale resolution than are immediately required. This is already the practice in most museum digitization projects, where usually only lowerresolution copies are released to the general public.

From the perspective of document image analysis, most research issues of preservation are related to digitization.

7. CONCLUSION

Transient problems may remain unsolved for a surprisingly long time.