RLG DigiNews
RLG's online newsletter for digital imaging and preservation

FAQ

Handwriting Recognition for Historical Documents

Author: Richard Entlich

OCR (Optical Character Recognition) seems to be widely used for providing searchable indexes of printed texts that have been scanned. Is it possible to do a similar thing with handwritten manuscripts and correspondence?

OCR Background

OCR is used to generate machine-readable text from printed documents. These are generally legacy documents from before the electronic publishing era, but may also be printed documents for which the original machine-readable text was discarded or lost.

OCR of printed text is a well-developed technology that has steadily improved in accuracy and flexibility. Initially limited to interpretation of numerals printed with special fonts, current day OCR software can deal with a multitude of fonts, character sets, languages, and page attributes. For extremely clean and well-scanned documents, the resulting text may be good enough to use for direct display purposes. More commonly, the OCR is somewhat "dirty" (i.e., contains errors) but is still accurate enough to form the basis of a quite usable machine-searchable index. Accuracy rates of 99.5% and higher (at the character level) are achievable for good quality source documents.

Handwriting Recognition Background

The conversion of handwriting to machine-readable text is usually referred to as handwriting recognition (HR). Computer scientists recognize two distinct classes of handwriting recognition. The better known of these is **on-line HR**, a real-time process usually employing a special stylus and pressure sensitive tablet that allows the direction and order of the writer's strokes to be monitored while writing. First popularized by the Apple Newton MessagePad, on-line HR is now available on most PDAs (Personal Digital Assistants).

The process of converting an existing handwritten document into machine-readable text is called **off-line HR** and is more closely analogous to OCR. Off-line HR is a far more daunting computing task and, as a result, is not as mature a technology as either OCR or on-line HR. The reasons are not hard to fathom.

Unlike printed text (that is, machine produced type), handwriting is subject to almost infinite variation. Cursive writing, in particular, can easily defeat human attempts to interpret it, as anyone who has attempted to decipher a doctor's handwriting can attest. Machine interpretation relies on reducing the scanned image to

some kind of recognizable pattern. Patterns may be missed because of vague word boundaries, overlapping letters, and great variations in the slant, spacing and shape of letters. Such variations may be modest within the writings of a single author, but are tremendously magnified across multiple authors. Further hampering recognition, handwritten documents tend to be "noisier" than printed ones due to smudging, staining, stray marks, underlining, and cross-outs.

Thus, early work in off-line HR, like that in OCR, focused on small, simple character sets such as numerals. Even today, much research and development is focused on highly constrained tasks such as reading cities, states and zip codes on hand-addressed mail, interpreting the dollar amount line on bank checks, or deciphering business forms, such as tax returns.

Methods for Off-line Handwriting Recognition of Historical Documents

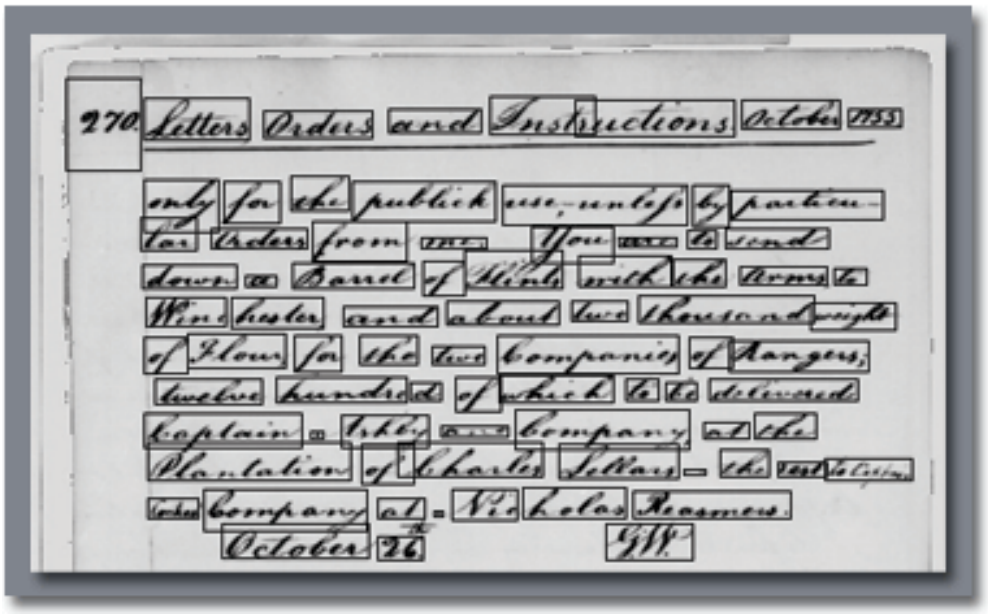


Figure 1. A portion of a scanned page from the Library of Congress's George Washington manuscripts. Rectangles have been drawn around where the words would be segmented. Also, dark lines which result from the scanning process have been removed from the sides. Note that the segmentation process is not perfect. "Winchester" in the fifth line and "Nicholas" in the next to last line have been divided into two parts. [1]

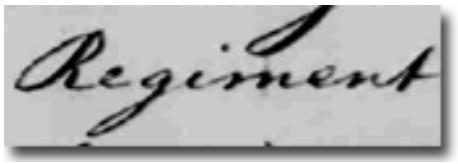
A small but steady stream of computer scientists has been trying to tackle the difficult task of deciphering cursive handwriting. The desire to improve access to large collections of important historical manuscripts has motivated most of this work. Scanned versions of the papers of Isaac Newton, U.S. presidents (especially George Washington), and the Archives of the Indies in Seville, Spain, to name a few, have served as recent experimental fodder.

In most cases, the objective of these experiments is less ambitious than full machine translation of handwriting. Instead, the goal is usually to recognize a subset of the most commonly used vocabulary (anywhere from a few hundred to one or two thousand words), usually within the writings of a single author. That vocabulary then serves as an index to support text queries. Limitations on vocabulary and authorship are intended to simplify the computational task so it can be done in a reasonable period of time, at an acceptable cost, and with a usable degree of accuracy.

Here are descriptions of a few of the different techniques being investigated:

Character segmentation attempts to identify individual characters and build them into words. This is exceedingly difficult to do with any degree of accuracy.

Word segmentation attempts to detect word boundaries, often supplemented by other document cleaning and filtering operations such as artifact removal, normalization of slant, smoothing, and binarization (converting grayscale images to bitonal). An effort can then be made to recognize the pattern made by an entire word and convert it to machine readable form without trying to identify individual characters.



Original grayscale image

Binarized with artifacts removed

With slant correction

Figure 2. Processing and normalization steps on a segmented word image prior to image matching. [2]

Word spotting is a form of off-line HR using word segmentation. In word spotting, the segmented words are first normalized to minimize variation and then similar images, which hopefully represent the same word, are clustered together. These groupings are called equivalence classes. No machine interpretation is done, only image matching. The groups of matched words are then displayed to a human operator who provides the text equivalent. Figure 3 shows a simplified diagram of the word spotting process, though stop words like "the" and "that" would normally be discarded. A subset of the most frequently occurring remaining words is used to create an index of the document.

Word spotting has also been applied in multiple author environments where word segmentation is not feasible, using different image matching techniques. (View enlarged image)

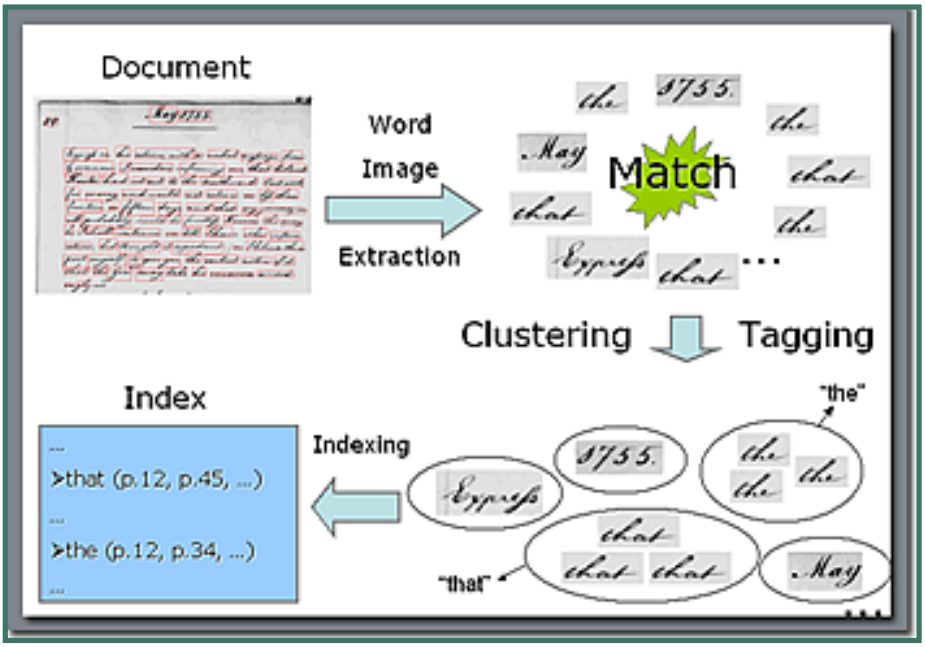


Figure 3. A conceptual diagram of the word spotting technique for indexing of matched word images. [3] (View enlarged image)

Statistical methods built on word segmentation are also being explored. Within a set of documents by a single author, a training subset is word segmented and manually transcribed. The images of the words are described using a highly formalized language based on the features (size, sequence of hills and valleys, etc.) of the particular image. The statistical correlation of the transcribed words with their feature-based descriptions is recorded for the entire set of training documents. Subsequently, a textual query can be made against a set of documents from the same collection that have been word segmented and feature described, but not transcribed. The query returns a set of word images (within a single line of the original document) most likely to match the query terms.

Transcript mapping is a technique used when a transcript of a handwritten document has been created, but it is unknown how the transcript corresponds to the location of words (pages, lines, and line position) in the original document. The existence of a transcript defines the vocabulary of the document, leaving the still non-trivial task of determining precisely where those words occurred.

Commentary

The amount of research activity and the variety of clever techniques being utilized in off-line HR should be gratifying for the archivists who maintain, and the scholars who utilize, handwritten historical documents. However, it should be noted that none of the work described here appears ready to emerge from the laboratory anytime soon.

Unconstrained machine translation of handwriting appears particularly far off, and may be unachievable. Even a less ambitious goal, such as software to reliably create partial indexes from good quality single author material, is unlikely to be met within the next several years.

However, enough progress seems to have been made for librarians, archivists, and scholars to become more involved in the ongoing research. Until now, there appears to have been little participation by those parties other than to provide sample documents and, on occasion, to serve on advisory boards.

For librarians and archivists, the future potential for machine translation should at least be considered when handwritten historical documents are digitized, particularly large collections by authors with legible handwriting. Since documents deemed worthy of digitization are likely to be of greater than usual significance, they are also good candidates for transcription and/or indexing. Accurate off-line HR depends on

scans with minimal noise and artifacts, so some additional effort to create very clean scans may be merited.

For those documents that are deemed so significant that it is worth fully transcribing them manually, the transcripts should record page, line, and word position to facilitate the potential to create indexes that can pinpoint the search term's location in the scanned document. (This presumes the document can be word-segmented, so the nature of the author's handwriting is again a consideration.)

Archivists could also advise computer scientists about how best to produce indexes that would interoperate smoothly with existing machine-readable finding aid standards such as EAD (Encoded Archival Description).

From the computer science side, more consultation with archivists and librarians familiar with the scanning of historical documents could avoid certain costly mistakes. For example, some of the researchers spent time cleaning up highly compressed JPEG files that suffered from severe artifacting around the text, instead of starting out with uncompressed or losslessly-compressed TIFFs.

Others have worked with low-resolution grayscale images that they have binarized using static thresholding techniques (that is, a single threshold value was used to binarize an entire page or collection of pages). Historical documents are usually scanned at 8-bit grayscale because they tend to be too tonally rich for satisfactory bitonal capture. However, some of the computer scientists seemed unaware of the availability of scanning software capable of dynamic thresholding and automatic background detection and suppression. Such software can produce bitonal scans with uniform contrast and text legibility even from originals with stains, fading, and uneven ink density.

In the meantime, if it is discovered that certain scanning practices would substantially improve the prospect for usable HR of historical documents, those standards should be promulgated to libraries and archives for consideration.

Finally, it is unclear what role scholars are playing in the development of systems for HR on their behalf. Though most of the details of crafting a successful off-line HR system fall within computer science and closely related realms, there are certain questions that only the end users of historical documents are in a good position to answer.

What corpora of historical documents would benefit most from being made searchable? If a search vocabulary has to be whittled down in size in order to reduce the computational load, which terms should be given priority for retention? Should the most common terms be kept, or should personal names, place names, or dates be preferred? What degree of inaccuracy can be tolerated before an index loses its value?

Conclusion

There is as yet no commercial or open source software for automatic transcription of, or the creation of searchable indexes from, handwritten historical documents. However, it is an active area of research and progress is being made. Continued advancement depends on the availability of funding. Librarians, archivists, and scholars may be able to push the agenda more effectively by partnering with computer scientists who share an interest in solving this challenging problem and improving access to significant historical archives.

Further Reading

Note: Much of the literature of off-line HR is highly technical. Some of the following papers provide a general overview of the subject, while others are best read for their abstracts, introductions, and conclusions (unless, of course, hidden Markov models and affine transforms are your cup of tea). All documents are PDFs.

Kane, Shaun, Andrew Lehman, Elizabeth Partridge, "Indexing George Washington's Handwritten Manuscripts: A Study of Word Matching Techniques." Technical Report of the Center for Intelligent Information Retrieval, University of Massachusetts, 2001.

Keaton, Patricia, Hayit Greenspan and Rodney Goodman, "Keyword Spotting for Cursive Document Retrieval," Proceedings of the IEEE Workshop on Document Image Analysis (DIA '97), in conjunction with the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '97), June 1997, San Juan, Puerto Rico, pp. 74-81.

Koerich, A. L., R. Sabourin, C. Y. Suen, "Large Vocabulary Off-Line Handwriting Recognition: A Survey," *Pattern Analysis and Applications*, v. 6, no. 2, pp. 97-121, July 2003.

Manmatha, R., "Word Spotting: Indexing Handwritten Manuscripts," DLI2/IMLS/NSDL Principal Investigators Meeting, Portland, Oregon, July 17-18, 2002.

Plamondon, Rejea and Sargur N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 22, no. 1, January 2000.

Rath, Toni M., Victor Lavrenko and R. Manmatha, "A Statistical Approach to Retrieving Historical Manusript Images without Recognition." Technical Report of the Center for Intelligent Information Retrieval, University of Massachusetts, 2003.

Tomai, Catalin I., Bin Zhang and Venu Govindaraju, "Transcript mapping for Historic Handwritten Document Images," Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02), pp. 413-418, September, 2002.

Verma, B., M. Blumenstein & S. Kulkarni "Recent Achievements In Off-Line Handwriting Recognition Systems," International Conference on Computational Intelligence and Multimedia Applications (ICCIMA '98), Melbourne, Australia, pp. 27-33, 1998.

Notes

[1] Image courtesy of R. Manmatha, Center for Intelligent Information Retrieval, University of Massachusetts. (back)

[2] Originally published in Rath, T.M., S. Kane, A. Lehman, E. Partridge and R. Manmatha, "Indexing for a Digital Library of George Washington's Manuscripts: A Study of Word Matching Techniques," Technical Report of the Center for Intelligent Information Retrieval, University of Massachusetts, 2002. Used with permission. (back)

[3] Adapted from Manmatha, R., "Word Spotting: Indexing Handwritten Manuscripts," DLI2/IMLS/NSDL Principal Investigators Meeting, Portland, Oregon, July 17-18, 2002. Used with permission.(back)

