Towards Searchable Indexes for Handwritten Documents

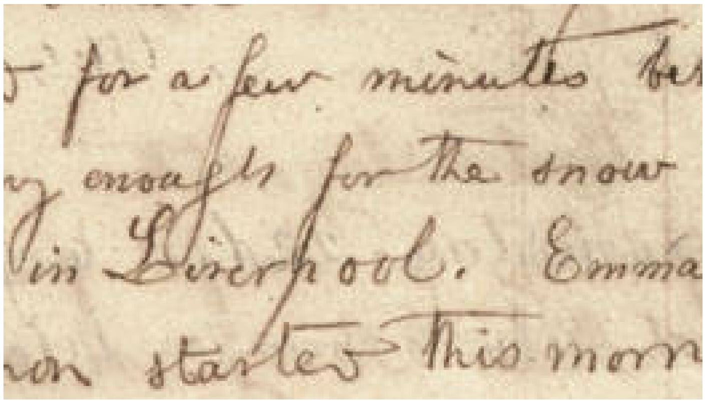
Douglas J. Kennard and William A. Barrett
Computer Science Department, Brigham Young University
{kennard, barrett}@cs.byu.edu

Abstract

While reliable, automatic transcription and indexing of handwritten historical documents remain long-term goals that have not been achieved, slow (but steady) progress continues to be made toward these goals. In this presentation, we provide a brief description of some of the difficulties encountered in handwriting recognition, as well as an overview of some approaches that have been taken in the past. We describe recent advances that show significant promise for creating search functionality for handwritten text. We demonstrate results from research we have recently done here at BYU to improve the quality of the input to recognition and indexing systems, and discuss some possible future work. We conclude that, although many of the problems with recognition and indexing of handwriting remain to be solved, the outlook is optimistic. While we do not anticipate reliable transcription, nor complete indexing within the near future, we do believe that useful, partial indexes and searches are within reasonable grasp.

Difficulties in Handwriting Recognition

Unlike the relatively clear and predictable size, spacing, and shape of machine-printed letters, handwriting is inconsistent and much less predictable. Different people write very differently, and even a single person usually exhibits a great deal of variation in the way he or she writes. As seen in Figure 1, the spacing between letters and words is not consistent, and parts of words may even overlap. The same letter can be shaped quite differently from one



s for a few minutes be
y enough for the snow
in Liverpool. Emma
on started this morn

Figure 1. A section of a hand-written page from the "Trails of Hope: Overland Diaries and Letters, 1846-1869" online collection available on the BYU Harold B. Lee Library website.

occurrence to the next. In addition, some letters may look very much like other letters (n's, m's, and r's, for example). Unlike "online" handwriting recognition for hand-held devices, we do not have temporal information or stroke order available to us, and are limited to the image, itself for recognition. Even though our human brains are very good at figuring out what is written, it is difficult to get a computer to reliably handle so much variation and ambiguity. The difficulties are even greater, of course, when we are dealing with historical documents, because of degraded or damaged paper, smeared or faded ink, bleed-through, and film or digitization artifacts.

Some Previous Approaches

We refer interested parties to the literature [1,2] for detail, but briefly discuss approaches using:

- Dynamic Programming
- Hidden Markov Models
- Human Reading Models
- Holistic (Word-level) Features

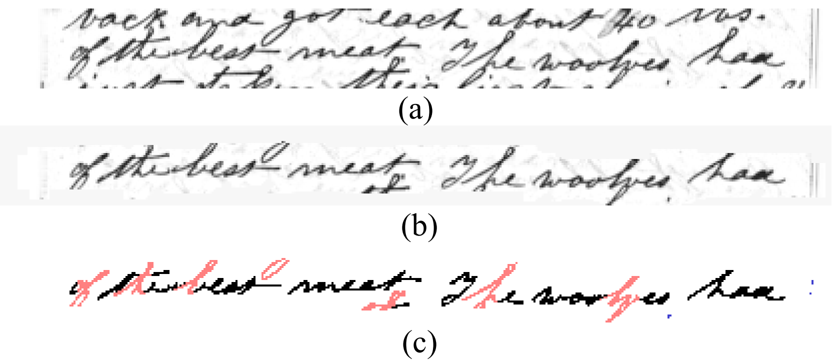
Recent Advances Elsewhere

In [3], a search engine for a collection of George Washington's manuscripts is automatically created using cross-lingual retrieval techniques, and made available at http://ciir.cs.umass.edu/irdemo/hw-demo/demo_intro.html as an online demo. The user enters a word to search for, and images of the most likely matches are returned in context. The paper and demo show that creating systems to search handwriting is not a task that is completely out of reach. However, the penmanship in the documents used is quite good, and 100 pages of the collection are manually transcribed as training examples for the system before it can be applied to the 987 test pages used in the demo.

In [4], documents are automatically "annotated" (indexed), and then the users of the index and images can correct errors as they come across them. An online demo of this is also available, at http://imadoc-ar.irisa.fr (in French).

Recent Advances at BYU

The accuracy and robustness of recognition and indexing systems depends a great deal on how well page segmentation works to provide the recognition/indexing system with input. While most handwriting we have seen can be split into lines of text by using profile methods, some handwritten lines are not straight, and require more advanced methods to split them. We have developed a method of splitting handwritten documents into images of individual lines of text using a binarized foreground/background transition count map and a min-cut/max-flow graph cutting algorithm. Early results with our method show that it will work in some difficult cases in which profile-based methods do not work properly. In addition to providing images of the lines of text, our method removes stray marks from other lines of text and flags marks that may be ascenders/descenders from other lines as ambiguous, as shown in Figure 2. This would allow recognition and indexing systems to take the ambiguity into account.



(a)

(b)

(c)

Figure 2. An example of our line segmentation method showing the line image for the middle line. a) Original image, including the line above and below. b) Grayscale segmented line image. c) Binarized version of the line image with ambiguous portions highlighted.

Future Work

As mentioned earlier, 100 pages of the George Washington manuscripts were manually transcribed as training data for the search engine in [3]. In general, this is typical of handwriting recognition systems—they require a large amount of training data. We discuss some ideas for trying to reduce the amount of initial training required before systems can begin to recognize or index automatically.

Conclusion

Because of the difficulties involved in handwriting recognition and automatic indexing, it is unlikely that we can automatically create full, flawless indexes in the near future. However, even partial indexes with errors will be more useful than no indexes at all. Recent research demonstrates that it is possible to create search engines for handwritten text, at least for good, consistent handwriting. By focusing our efforts on reducing the amount of manual training required for recognition and indexing systems, we believe that use of such systems will be more feasible. While these systems will not be perfect, interfaces provided to the end-users can allow them to correct errors as they are discovered, making future searches by others more accurate.

References:

[1] Vinciarelli, A. “A survey on off-line Cursive Word Recognition,” *Pattern Recognition,* 35 (2002) pp. 1433-1446.

[2] Koerich, A. L., Sabourin, R., Suen, C. Y. “Large vocabulary off-line handwriting recognition: A survey,” *Pattern Analysis and Applications*, 6 (2003) pp. 91-121.

[3] Rath, T. M., Manmatha, R., Lavrenko, V. “A Search Engine for Historical Manuscript Images,” *SIGIR'04*, Jul. 25-29, 2004, Sheffield, South Yorkshire, UK.

[4] Coüasnon, B., Camillerapp, J., and Leplumey, I. “Making Handwritten Archives Documents accessible to Public with a Generic System of Document Image Analysis,” *Int'l Workshop on Document Image Analysis for Libraries*, Jan. 23-24, 2004, Palo Alto, California, pp. 270-277.