# Progress with Searchable Indexes for Handwritten Documents

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

**Abstract**

Automatic transcription and indexing of handwritten historical documents remains a long-term goal that will probably not be achieved in the near future. However, the simpler problem of searching handwritten documents using word spotting techniques is well within grasp. We report on our current progress toward creating such searchable indexes for handwritten documents, as well as planned directions for future work. We also demonstrate some applications in which such an approach could be put to immediate use in aiding family history and genealogical research.

## 1. Introduction

At the Family History Technology Workshop during 2006, we discussed some of the difficulties involved in offline handwriting recognition (HR), and particularly when dealing with historical documents. We briefly described some general HR approaches used by previous researchers, and also referred to work done elsewhere[1], in which Google™-like searches for a collection of George Washington manuscripts are demonstrated. We reported our own early work toward such an end, which included some image preprocessing and a novel method for handwritten text line segmentation.

In this paper, we report on our current progress and continued efforts toward making handwritten documents searchable. Specifically, we describe our efforts to:

1) Automate word separation
2) Reduce the amount of training required through interactive training
3) Provide a usable search system similar to the George Washington demo

In addition, we describe some applications (other than search-engine-like applications) in which word-spotting techniques could be put to use immediately upon completion. These include prioritizing extraction redundancy and arbitration, and aiding extraction (indexing) within unstructured handwritten records.

## 2. Automatic Word Separation

The accuracy of both handwriting recognition and word spotting depends on how well handwritten words can be located and separated from each other on the handwritten document page. We are currently experimenting with methods of automatically finding and separating words, and we give a high-level report on our current progress and results, while emphasizing the fact that this research is a work in progress. Our completed word separation method will be entered in a competition later this year against other word separation methods. The results of the competition will be announced and reported at the International Conference on Document Analysis and Recognition (ICDAR 2007).

## 3. Interactive Training for Training Set Reduction

In order for handwriting recognition or word spotting to work, systems must first be given training data, which consists of examples of words that are contained in the documents along with the labels (transcriptions) of those example words. Due to the great variability in handwriting, a large amount of training data is usually necessary, even for documents written by a single author.

We describe our proof-of-concept interactive training application (see Figure 1), in which the person labeling the training data manually enters word labels as they are highlighted. At first, the application highlights all words, but as it comes to words that appear to be very similar to words that have already been labeled, those words are skipped. The idea is that since many words are used frequently in the English language (and most other languages), less redundant training occurs, resulting in a net reduction in the amount of training that is needed for a given level of recognition accuracy.

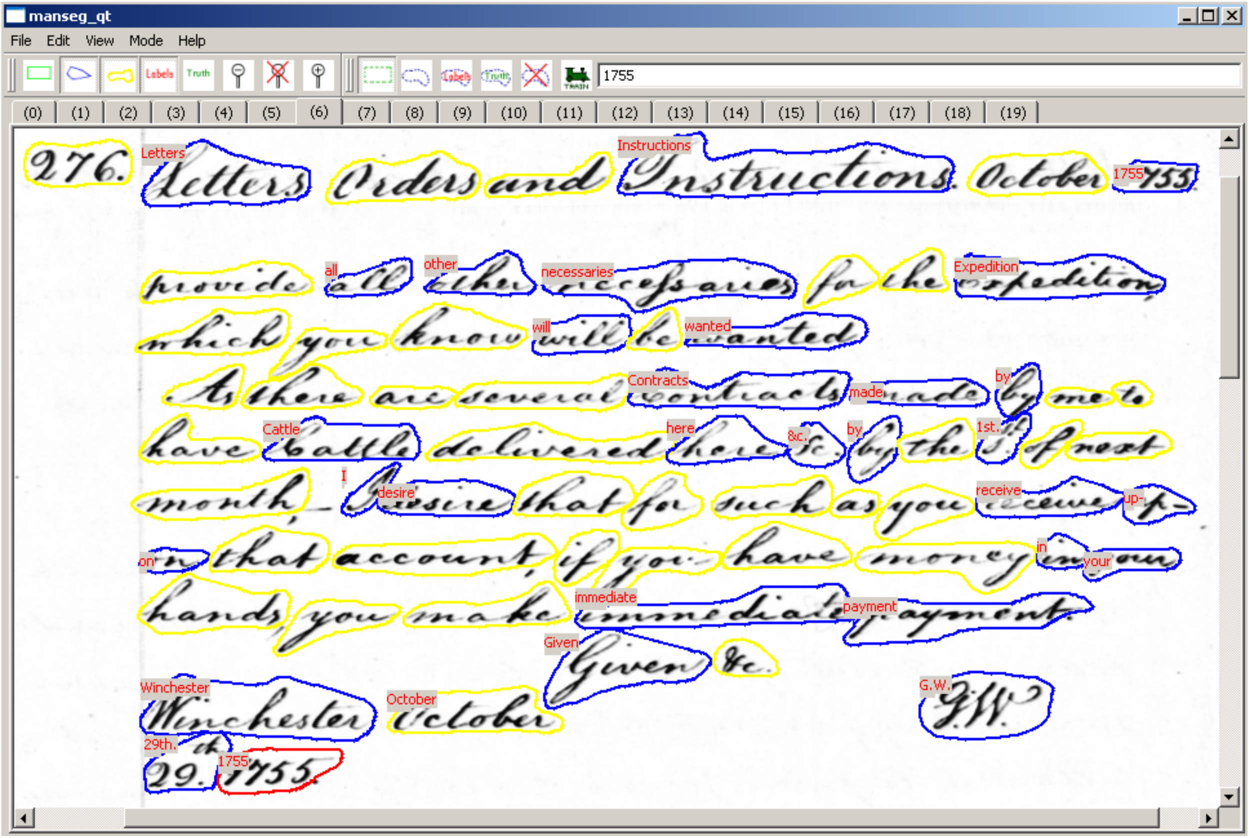

Figure 1. Interactive training application. The user is required to label only those words (dark blue) that are sufficiently different from any words that have already been labeled. Words highlighted in light yellow are automatically skipped. The document is from the Washington collection made available by Rath and Manmatha[3].

Our preliminary experiments with interactive training, published in [2], are encouraging. Our experiments are conducted using a collection of 20 pages of the George Washington manuscripts made available by Rath and Manmatha[3]., as well as a collection of 30 pages from Jennie Leavitt Smith's diary, downloaded from the "Mormon Missionary Diaries" online collection of the BYU Harold B. Lee Library, available at http://www.lib.byu.edu/dlib/mmd/. Training is done on at most 1,000 words from each collection, starting at the beginning, and does not overlap the test set. For the Washington manuscripts, the test set consists of 2,368 words, and for the Smith Diary, the test set consists of 1,791 words. Results of interactive training are compared with results of non-interactive sequential training. The recognition ratio (defined as the number of words in the test set that get labeled correctly, divided by the total number of words in the test set ) is computed by comparing the test results of either method with ground truth data that is entered manually.

In Figure 2, the graph shows a comparison of the recognition ratio (vertical axis) at various amounts of training (horizontal axis) for the Washington manuscript test. The recognition ratio is graphed for both interactive training and for labeling words sequentially without interactively skipping similar words. As seen in the graph, less training is required when using interactive training than when using non-interactive, sequential training to achieve the same recognition ratio. For example, to achieve a recognition ratio of 0.49, our method reduces the amount of training from about 900 words to about 700 words, which is more than a 22% reduction.

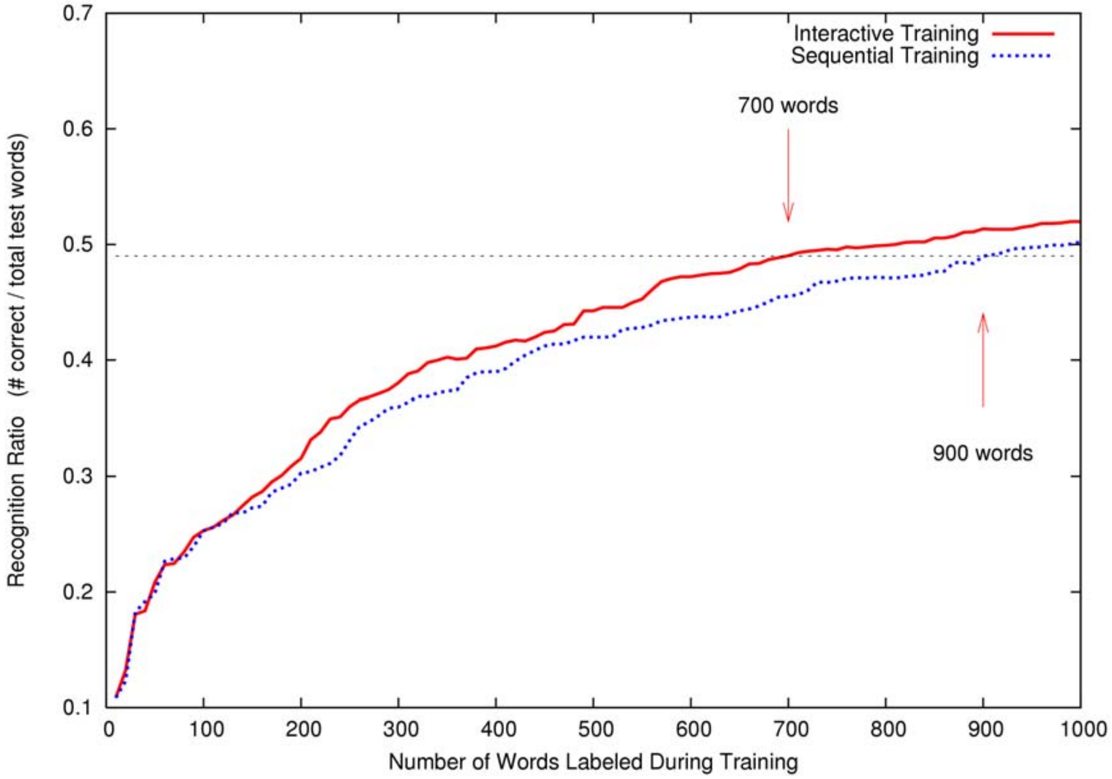

Figure 2. Comparison of interactive training vs. sequential training for Washington manuscript test. Less training is required to reach the same recognition ratio when interactive training is used.

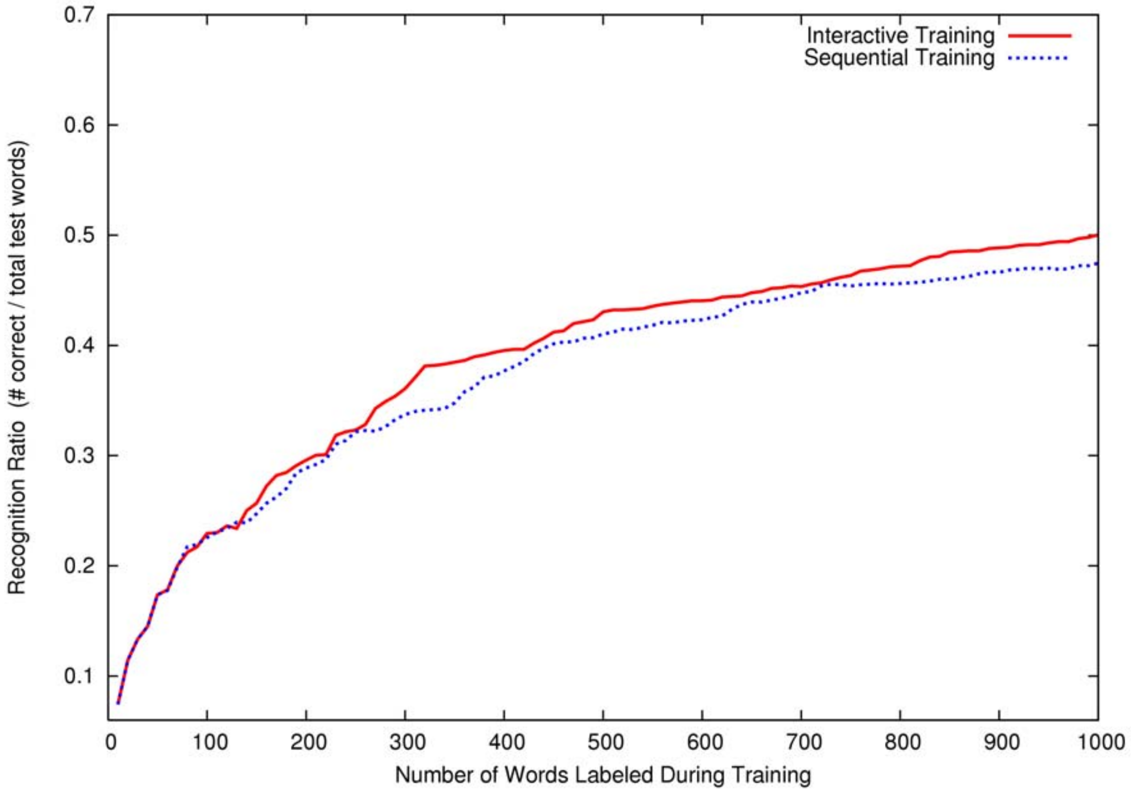

Figure 3. Comparison of interactive training vs. sequential training for Smith diary test. Slight improvement is observed when interactive training is used.

Slight reduction is observed in the amount of training required when using interactive training for the Smith diary test, as well, although the improvement is less dramatic than the improvement with the Washington manuscripts (see Figure 3). These two preliminary tests support the idea that less training is required when training interactively in the manner described instead of training on all words sequentially, given the assumption that excellent word separation is provided and a similarity threshold parameter is well-chosen.

## 4. Search System Demo

We report on the status of our efforts to implement a search engine demo similar to the George Washington search demo described in [1], which is available to the public at http://ciir.cs.umass.edu/irdemo/hw-demo/demo_intro.html.

Our initial demo builds off of the work of the authors of [1], and provides a baseline method to which we can compare our future work. The demo system also provides a framework within which we can test new ideas and demonstrate the effectiveness of those ideas as our work progresses.

## 5. Other Applications

In addition to search engines for handwritten documents, there are other useful applications in which the same algorithms and technology could be used. We suggest two:

> -Prioritizing Extraction Redundancy and Arbitration: Use recognition / word-matching confidence levels to determine which extracted information should be entered by a second extractor in the near-term, and which is probably correct and can wait for redundant extraction and arbitration.
>
> -Aiding Extraction within Unstructured Records: Many records are handwritten and unstructured, but have the same information in each form. Recognition / word spotting could be used to find words of interest, such "born," "died," "mother," etc. and provide highlights similar to the highlights that aid users in the fields of structured forms in current extraction projects.

## 6. Conclusion

While robust automatic handwriting transcription is probably not achievable in the near future, we report progress toward other more easily attainable goals, such as searching within handwritten documents. Such applications are within reach of current technologies. We also suggest two related applications for which the same technologies could be used.

## References:

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

[2] Kennard, D. J., Barrett, W. A. "Interactive Training for Handwriting Recognition in Historical Document Collections," *DRR'07*, Jan. 28 - Feb. 1, 2007, San Jose, CA, USA.

[3] Rath, T. M., Manmatha, R. "Word Image Matching Using Dynamic Time Warping," *CVPR'03*, Jun. 2003, Madison, WI, USA.