Green ICR: Semi-Automated Census Record Indexing with Emphasis on Human Computer Interaction

Robert Clawson, William Barrett

Brigham Young University, Provo, UT, USA

ABSTRACT

Human-based computation is an approach that utilizes the abilities and strengths of both humans and computers to achieve a symbiotic interaction that is stronger than either agent in isolation. We propose a system that amplifies the capacity of a human indexer by adding an intelligent handwriting recognition engine to the indexing process. This recognition engine will learn patterns in handwriting as the indexer works, and then will amplify the indexer by automatically labeling handwriting with similar patterns. The recognition engine may also prompt the user to label examples that will best help it to learn. Preliminary results show that applying handwriting recognition technology could significantly reduce the number of fields an indexer is required to hand label. As a byproduct, we also believe the proposed system will be more interactive and more enjoyable for the indexer.

1. INTRODUCTION

Currently, thousands of volunteer hours are dedicated to indexing historical documents. An index allows researchers to quickly find an item of interest, rather than manually searching through hundreds of records a line at a time. FamilySearch indexing is an ambitious crowd sourcing project where volunteers index historical records one field at a time. They record information such as name, age, gender, marital status, and place of birth. These indexes allow genealogists to find information about ancestors contained in many types of documents with simple queries. FamilySearch's vast and growing collection of indexed records is possible only through the combined labor of the tens of thousands of volunteers who give hours of time to the effort.

Repetitive tasks, like indexing, are strong candidates for automation. Thousands of man hours could be saved, and more work accomplished, if indexing could somehow be automated. Unfortunately, there are many image processing challenges to overcome in an end-to-end automated indexing system, and one of the most challenging is automated handwriting recognition. Recognizing handwriting automatically has been studied for many years, and yet continues to be impracticable except in constrained circumstances.

Rather than apply a monolithic, fully automated recognition to the task, we investigate the practicability of a semi-automated learning system which uses both the strengths of human indexers and handwritten word recognition technology.

2. RELATED WORK

A2ia has published work on an automated system for indexing French census registers.[1] Their work is industry leading in the field of automated indexing. However, the system is built to fit the paradigm of a recognition engine doing the best it can without human interaction. The research proposed in this paper should prove to be more robust, more widely applicable, and yield more accurate results.

More related to the proposed research are recent papers on semi-supervised labeling of historical weather reports and Lampung characters.[2,3] These papers more closely approaches the problem addressed in the research proposed in this paper, but are still focused more on the recognition engine than on how to interact and learn from the user in real time.

Though in a different domain, the key contribution in Intelligent Scissors[4] is an excellent metaphor for the intended contribution of the proposed research. Though graph search methods had been in use for many years in image segmentation, the addition of a human providing simple guidance in Intelligent Scissors made human guided segmentation a reality as a real time tool. Likewise, though handwriting recognition has been studied for many years, by pairing it with real time human guidance, an efficient and improved indexing system is the intended result of this proposed research.

Finally, in relation to the preprocessing steps mentioned in this paper, more details have been published and are available.[5]

3. THESIS STATEMENT

We claim that linking indexer and computer in a tightly coupled, symbiotic relationship for the task of indexing will provide (a) a more efficient, accurate, and reproducible system, (b) a framework for incremental learning for both user and computer, (c) a domain-constrained system that improves and is adaptable, (d) possibly a more engaging experience for the user.

4. RESEARCH DESCRIPTION

This section will summarize the different algorithms that we propose to use to substantiate our thesis statement. We will be constraining our work to documents with a tabular structure. While the concepts of the research will extend to other document types, tabular documents are common in the domain we are targeting and are the most direct application of already available research. Tests will be run on the 1920 Utah and Delaware census collections. These collections have thousands of pages, with tens of thousands of names. They are already indexed, and provide a wealth of training data.

4.1 Preprocessing

Document images in a collection are scaled, translated, binarized, zoned, and handwriting slant is removed. A template for the collection is used in these processes. After preprocessing, each field in the table can be isolated. This is an absolute requirement; without the ability to isolate handwriting samples in the document, there is no way to compare them to each other and suggest labels. The preprocessing step is nontrivial, but also is not the focus of the proposed research, so existing technology will be used ”out of the box”. These preprocessing steps have been demonstrated to work reliably on available datasets, but ultimately some recognition error will be attributable to preprocessing.

4.2 Initialization

The base case for this system is a fresh collection of tabular document images, with or without labels. In this situation, the indexer simply begins manually labeling fields, not unlike the current paradigm in FamilySearch indexing. However, the indexer will be encouraged to work down columns, rather than across rows. As the fields in a column begin to be populated, the computer agent will begin comparing unlabeled fields to labeled fields. When a confidence threshold is crossed it will begin to automatically label fields.

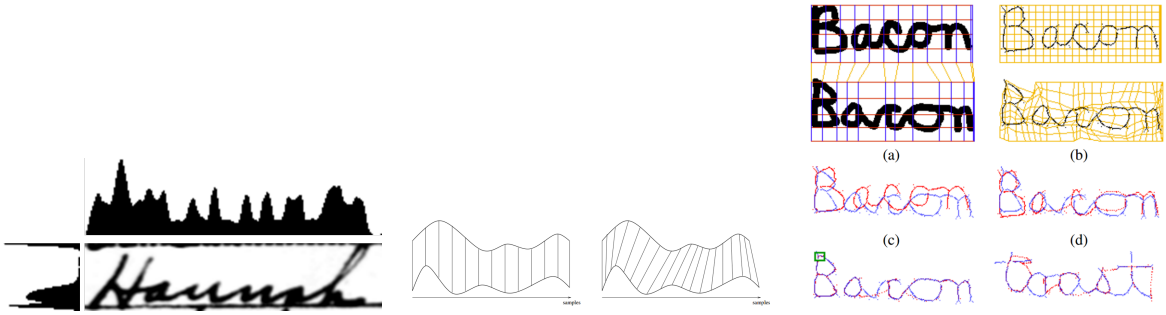
4.3 Word Matching

Generating a good similarity score between word images is central to the learning algorithm’s accuracy. This can also quickly become the bottleneck computationally of the system. We propose three metrics of increasing discrimination and algorithmic complexity that will be used to match word images (see Figure 1). The first is a simple correlation of the word image profiles. This method is fast, and may prove to be sufficient for fields with a limited vocabulary. The second is the method proposed by Rath and Manmatha, where word image profiles

are matched using dynamic programming.[6] Finally, Kennard's word warping algorithm will be used for the most difficult fields.[7]

Some experimentation will be needed to decide when to use the three different word matching algorithms. A simple heuristic would be to switch at some threshold of the number of different word types in a column. As long as the training set is limited in size, any of the three should be feasible at interactive speeds. However, in the interest of being able to scale to larger datasets and also to save compute cycles when a simpler algorithm would be sufficient, these three metrics are all presented.

As indexers process through a collection of records, there will be a growing number of samples to compare to. A simple k-NN classifier will be used to choose the best label.



(a) Profile Correlation (b) Profile Matching with Dynamic Programming (c) Word Morphing

Figure 1. Three metrics for word matching.[6,7]

4.4 Training

There is a question of how much data to use as labeled examples. This data can come from recently labeled examples from the same author, and also from other sources and other authors. The examples from the current author are most likely to yield accurate results, but are few in number. On the other hand, the vast collection of previously seen examples of a given word may be written very differently, and results tend to indicate much lower performance under multiple authorship.[5] For this reason, we will at least begin by including only examples from the current author in the training set. This will be reasonable for cases where there is enough data from each author to train and begin to reap benefits before the author changes.

4.5 Human Computer Interaction

We have several ideas in designing the interaction between the indexer and the computer agent. First, the computer may interrupt the indexer and ask that a particular field be labeled. This would require some sort of active learning heuristic to decide which remaining unlabeled example, if labeled, would best discriminate between all unlabeled examples (always in the context of the current author). Second, the computer agent can cluster the remaining unlabeled examples in the column of focus and highlight in different colors the fields on the page according to cluster (see Figure 2). Then, the indexer can type the label only once, and it will apply to all fields in a cluster. The indexer could also quickly correct any mistakes in the computer agent's labeling. Finally, the computer could potentially flag errors that it suspects have been made by the indexer. Careless mistakes like misspelling or inputting a label for the wrong field could be corrected.

5. RESULTS

Although this paper focuses on proposed future work, some preliminary results exist. Readers are referred to another paper for classification accuracies under several categories using word morphing in the context of census record indexing.[5]

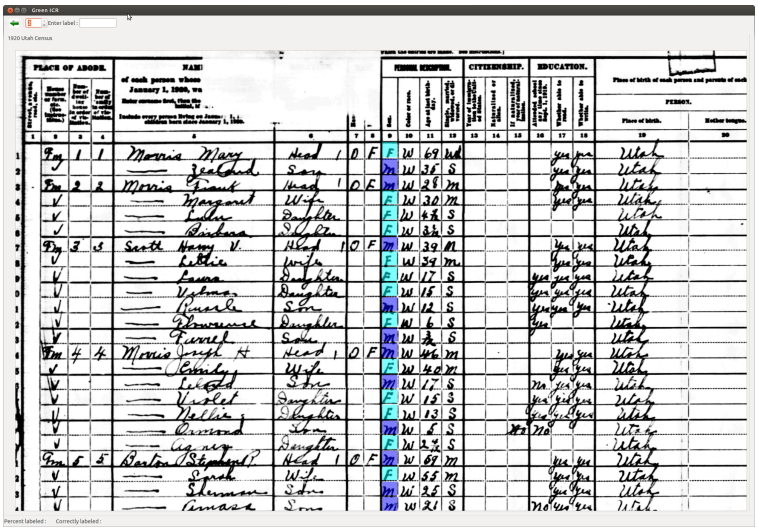


Figure 2. Possible model for a user interface where indexer interacts directly with the image, and the computer can cluster words it has found to be similar (best viewed in color).

6. CONCLUSION

There is great potential for future work beyond what is proposed here. Exploiting relationships between fields in a row of data is a simple example (if a person is a wife, they are probably married and female as well). Validation of the proposed system will be in terms of efficiency, and accuracy. We expect a dramatic increase in indexing throughput, as well as increased accuracy.

REFERENCES

[1] Sibade, C., Retornaz, T., Nion, T., Lerallut, R., and Kermorvant, C., "Automatic indexing of french handwritten census registers for probate geneaology," in [*Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*], 51–58, ACM (2011).

[2] Richarz, J., Vajda, S., and Fink, G., "Towards semi-supervised transcription of handwritten historical weather reports," in [*Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*], 180–184, IEEE (2012).

[3] Vajda, S., Junaidi, A., and Fink, G., "A semi-supervised ensemble learning approach for character labeling with minimal human effort," in [*Document Analysis and Recognition (ICDAR), 2011 International Conference on*], 259–263, IEEE (2011).

[4] Mortensen, E. and Barrett, W., "Intelligent scissors for image composition," in [*Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*], 191–198, ACM (1995).

[5] Clawson, R., Bauer, K., Chidester, G., Pohontsch, M., Kennard, D., Ryu, J., and Barrett, W., "Automated recognition and extraction of tabular fields for the indexing of census records," in [*IS&T/SPIE Electronic Imaging*], 86580J–86580J, International Society for Optics and Photonics (2013).

[6] Rath, T. and Manmatha, R., "Word image matching using dynamic time warping," in [*Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*], **2**, II–521, IEEE (2003).

[7] Kennard, D. J., Barrett, W. A., and Sederberg, T. W., "Word warping for offline handwriting recognition," in [*Document Analysis and Recognition (ICDAR), 2011 International Conference on*], 1349–1353, IEEE (2011).

[8] Little, G. and Sun, Y., "Human ocr: Insights from a complex human computation process," in [*Workshop on Crowdsourcing and Human Computation, Services, Studies and Platforms, ACM CHI*], (2011).