Better Historical Document Indexing Using Waypointing and ROI Data

Kevin Bauer

Brigham Young University, Provo, UT, USA

ABSTRACT

Currently one of the biggest bottlenecks in the historical document indexing process is the amount of time needed to prepare a document for indexing. Although many technologies exist which can assist in this process, much of this work is still done by hand. This paper proposes an interactive system that assists the user in quickly preparing large sets of document images for indexing. Such a system would extract waypointing data from title cards, assist in clustering large groups of documents, and then perform layout analysis on the individual documents to further extract header fields, regions of interest (ROIs), and other metadata. While the system would automate the process as much as possible, user input would be incorporated to ensure a high degree of accuracy. Such an interactive system would be highly robust and accurate while amplifying the user′s efforts several times. The primary contribution of this work will be the creation of novel algorithms for collecting dataset-level metadata through the use of title cards and document headers. While much research has been done on document analysis, most of it has been focused on individual documents while ignoring valuable data that can be obtained by intelligently clustering and grouping large sets of similar documents. A further contribution will be the pairing of individual data fields with the ROI from which they were indexed. Because most indexing systems require input images to be prepared by hand, ROI data is usually not captured or is discarded. However, a system that makes it possible to include the ROI data has many potential benefits, especially as a means of leveraging mobile devices as tools for volunteer indexing. Finally, the proposed system could be used to create searchable, ground-truthed datasets that would benefit the document analysis community as a whole.

Keywords: Document Clustering, Layout Analysis, Waypointing

1. INTRODUCTION

In recent years the LDS Church has spearheaded a massive, worldwide program seeking to digitize and create a searchable index of every historical document they have available. Since the church′s Granite Mountain Vault contains over 18 petabytes of data, this is a monumental task with several steps, the most well-known of these being the task of using volunteers to manually index the important information in each document. Two of the most important steps, however, take place before indexing. First, the nature of the collection must be understood: how it is organized, bounded, and sorted, how genealogists will want to browse it, and what key data fields will eventually need to be extracted by indexers. This step is usually carried out by domain experts: highly trained genealogists with an understanding of the language, culture, and history associated with the record set. The next step is to examine the records as a stream of images and look for points where the record changes. These can be changes in date, location, or the layout of the document. These points of change are referred to as waypoints because they can be used to guide an individual in finding their way to specific information, and throughout this paper we′ll refer to the process as waypointing. Figure 1 gives an example of some common fields in a document header: the month, location, and year, with the points of change marked and underlined. As can be seen from the location field, sometimes values may be repeated (in this case Provo appears twice).

Waypointing can be seen as both an intermediate step towards full indexing of a collection as well as a means of making documents available for browsing without requiring them to be fully indexed. Despite these benefits, the limited availability of tools and volunteers has made waypointing a bottleneck in the indexing process.

In addition to the amount of time required for the process, waypointing is also limited in the amount of data that is extracted. In the context of FamilySearch waypointing is currently performed by a small group of volunteers looking at sets of images. To save time only document-level waypoints (such as those visible when

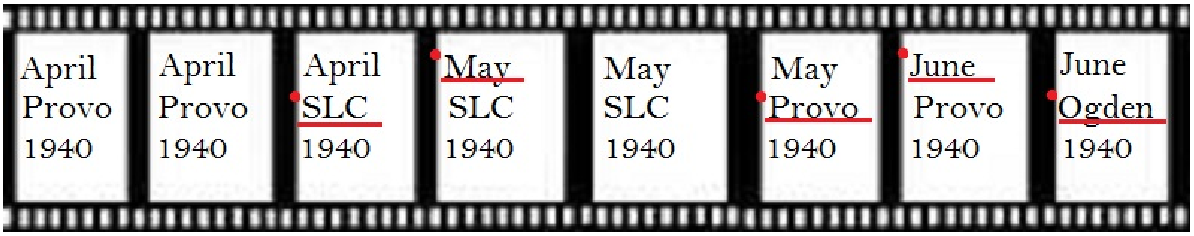


Figure 1. Points of change in a microfilm roll

examining a set of image thumbnails) are extracted. Field level waypoints such as changes in location, date, or author, are discarded, resulting in the loss of potentially relevant data.

This paper proposes a method for simplifying the waypointing process, allowing large document sets to be waypointed quickly without requiring a high degree of domain expertise. Document analysis methods that have not been previously applied to the problem of waypointing are also examined, and the creation of a user-guided waypointing system that incorporates these methods is proposed, with the goal of reducing the time spent waypointing large document sets while also increasing the number of field-level data that is captured.

2. PREVIOUS WORK

The field of document analysis is widely studied in the image processing community, and many algorithms exist for word spotting, zoning, and automatic structure analysis. Mao et al[1] present an algorithm that clusters documents by building trees to represent each document′s structure and minimizing the edit distance between these trees. The results they present are promising, but their algorithm deals primarily with digital born documents whose structure is fairly different from the scanned historical documents that are of interest to genealogists. Surdeanu et al[2] propose an unsupervised document clustering algorithm that is a hybrid of hierarchical clustering and expectation maximization (EM) clustering. Their method also relies on the document being digital and would require adaptation before it could be applied to the domain of historical document analysis.

Narrowing in on the process of historical document indexing in particular, Barrett et al[3] propose the creation of a Digital Microfilm Pipeline, an automated process that takes a set of documents from a roll of microfilm to a digitized, searchable index. The key elements of such a system are identified, including scanning, cropping, deskewing, registration, zoning, and OCR of machine-printed and handwritten fields. In recent years BYU researchers have developed some promising new techniques for registration[4,5] and handwriting recognition.[6] The process was recently revisited by Clawson et al.[5] However, these works did not address the problem of document clustering and waypointing in any great detail, and a system for addressing this problem has yet to be developed.

3. PROPOSED PROCESS

The first step in the process will be to automate image-level waypointing as much as possible, thus removing the need for the user to browse the documents as a stream of thumbnails. This can be done by automatically separating the different types of documents into clusters based on their type (title boards, census records, birth certificates, etc). Individual clusters will then be registered to a small number of randomly chosen images within their cluster. This will allow us to identify misclassified images by examining the sum-squared error. The user will then manually identify the misclassified images and place them in the correct clusters.

Once the images are clustered and registered the next step is to perform document-level waypointing. Next, areas on the page where important information can be found are identified. These areas are commonly referred to as Regions Of Interest (ROIs). In the context of title board images these ROIs identify every piece of information in the image. Since each title board contains information about several documents we propose that each of these ROIs and the information they contain be extracted and linked to the corresponding documents in each collection. Because title boards vary dramatically in structure this step will need to be performed manually, but will be worth the effort because of the high information density of each title board. Figure 2 gives an example of some of the information that can be extracted from a titleboard.

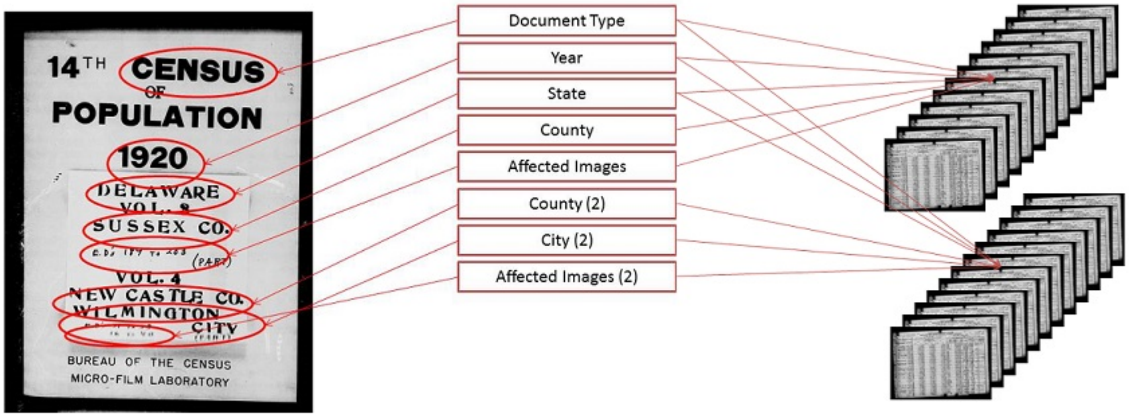


Figure 2. Extraction of information from a title board

In the context of tabular document images the ROIs include fields in the document header and footer as well as the individual cells in the table. The nice thing is that because we′ve clustered and registered the documents before this we need only find the regions of interest for one document per cluster, and then apply it uniformly to each of the other documents in that cluster.

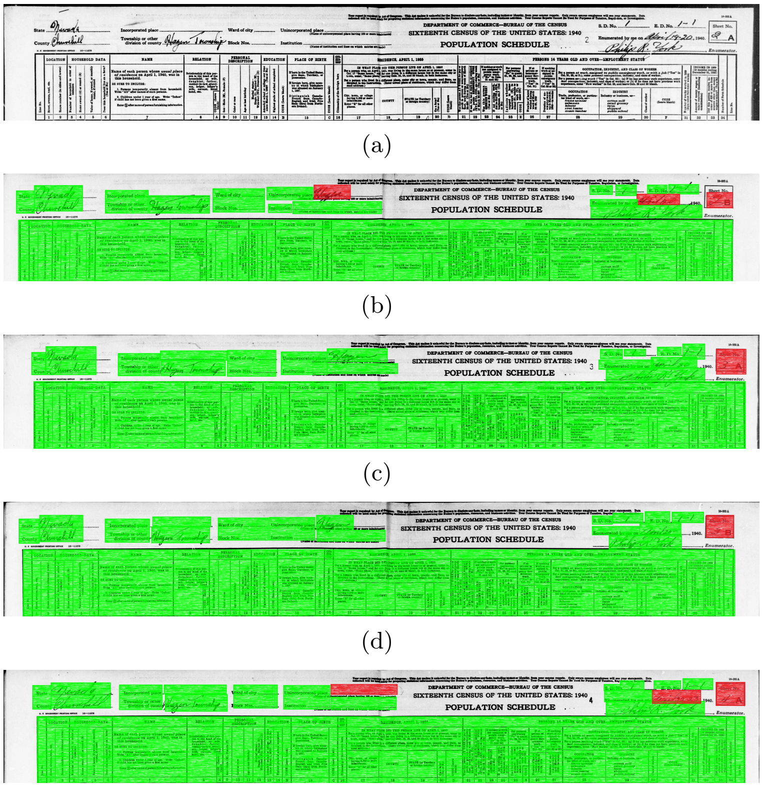
Another benefit of having registered the images beforehand is that it allows us to quickly capture all of the data in the document header. The user begins by manually recognizing and entering every piece of information in the header of the first document (this can be partially automated using the techniques described in[5]). Then the header of the next image is examined to see if any of the fields are different, with the user being prompted to manually enter these. Because the header contains mostly machine-printed or relatively static information such as the year, location, or author, these points of change will be relatively infrequent. Figure 3 shows the headers of 5 1940 census records that appeared next to each other on a microfilm roll. Fields that are the same as the previous image are marked in green, while points of change are marked in red.

Once the header information is captured the document clusters can be further refined based on any subset of the fields represented in the header (author, date, location, etc.) The resulting clusters of images can then be immediately made available for searching based on any of these fields. Further, all this data can be collected without requiring prior domain knowledge. Rather than starting from scratch, the domain experts will be handed a rich set of digitized searchable data, simplifying their task and reducing the amount of time they have to spend in preparing a document for indexing.

4. CONCLUSION

This paper proposes a user-guided waypointing and clustering process for historical document analysis. The need for refinements in current document clustering algorithms is identified, and a process for extracting information from title boards and document headers is suggested. The end product of this work will be an interactive GUI allowing a user to perform all these tasks on a set of documents in approximately the same amount of time it takes to waypoint the documents using current, less information-rich methods. If this tool is developed properly it could be general enough to accommodate any document without relying on prior domain knowledge it would allow domain experts to focus on the more difficult task of understanding the collection as a whole and deciding which information to extract. Ideally, this system would be of great benefit in this task as well by providing a rich set of data about the document.

Each step of this process has been designed with the digital microfilm pipeline in mind. By capturing all the ROIs and header data we have all the info we need to attempt to automatically index the document. Shortcomings in current handwriting recognition technology make this unfeasible, but current research into user-assisted or green OCR looks promising, and the proposed method would feed into such a system nicely.


DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS
SIXTEENTH CENSUS OF THE UNITED STATES: 1940
POPULATION SCHEDULE

(a)

(b)

(c)

(d)

(e)

Figure 3. Most of the header data does not change from one document to the next

One problem with the field of document analysis is that despite the growing amount of ground-truth data available most of the corpora used are ad-hoc datasets created to address a specific language and document type. The proposed system hopes to alleviate some of this by providing a means to quickly and effectively create such datasets, with the hope that larger, standardized datasets will emerge over time, allowing for a more meaningful comparison of document analysis algorithms and techniques and leading to improvements in the field.

REFERENCES

[1] Mao, S., Nie, L., and Thoma, G. R., "Unsupervised style classification of document page images," in [*IEEE International Conference on Image Processing*], 510–513, Accepted (2005).

[2] Surdeanu, M., Turmo, J., and Ageno, A., "A hybrid unsupervised approach for document clustering," in [*Proceedings of the Eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*], 685–690 (2005).

[3] Barrett, W. A., "Digital mountain: From granite archive to global access," in [*First International Workshop on Document Image Analysis for Libraries*], (2004).

[4] Hutchison, L. A. D. and Barrett, W. A., "Fast registration of tabular document images using the fourier-mellin transform," in [*Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL'04)*], *DIAL '04*, 253–, IEEE Computer Society, Washington, DC, USA (2004).

[5] Clawson, R., Bauer, K., Chidester, G., Pohontsch, M., Kennard, D., Ryu, J., and Barrett, W. A., "Automated recognition and extraction of tabular fields for the indexing of census records," in [*Document Recognition and Retrieval*], (2013).

[6] Kennard, D. J., Barrett, W. A., and Sederberg, T. W., "Word warping for offline handwriting recognition," in [*11th International Conference on Document Analysis and Recognition(ICDAR2011)*], **1**, 1349–1353 (2011).