Digitization and Search, A Non-Traditional Use of HPC

Liana Diesendruck, Luigi Marini, Rob Kooper, Mayank Kejriwal, Kenton McHenry
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
Email: {ldiesend, lmarini, kooper, kejriwa1, mchenry}@illinois.edu

We describe our efforts in developing an open source cyberinfrastructure to provide a form of automated search of handwritten content within large digitized document archives. Such collections are a treasure trove of data ranging from decades ago to as far as the present. The information contained in these collections is also very relevant to both researchers who might extract numerical or statistical data from such sources as well as the general public.

With the push to digitize our paper archives we are, however, faced with the fact that though these digital versions are easier to share, they are not trivially searchable as the digitization process produces image data and not text. This inability to find and/or identify contents within these collections makes this data largely unusable without a lengthy and costly manual transcription process carried out by human beings.

To carry out the search we build on top of a computer vision technique called word spotting. A form of content based image retrieval, it avoids the still difficult task of directly recognizing the text by allowing a user to search using a query image containing handwritten text and ranking a database of images in terms of those that contain more similar looking content. In order to make this search capability available on a large archive, three computationally expensive pre-processing steps are required, Figure 1. First, forms are segmented into individual units of handwritten information. In the case of the 1930 Census data collection, which contains approximately 3.6 million spreadsheet-like forms, this entails breaking the form images into sub-images of individual cells that contain the information about the individuals recorded in the Census. Second, the extracted sub-images are processed so as to extract features and descriptors that represent the handwritten contents within them. The utilized word spotting method results in a 30 dimensional vector derived from the frequency components of the darker ink pixels [1]. The distance between two such signature vectors can be used to determine how similar the handwritten contents of their cell sub-images are. Third, an indexing step organizes these extracted signatures into a binary tree structure to enable fast user queries. For the 1930 Census data this involves organizing nearly 7 billion sub-images using a hierarchical agglomerative clustering. Organizing the entire collection at once isn't practical, thus we instead break this step into multiple index construction steps based on states, categories, and microfilm reels passing the cost of a longer search time on to the user in the event that multiple states or categories are to be search simultaneously.

Complementing the computer vision based indexing, we also design the framework around a passive crowd-sourcing element. In order to accommodate the required query image for the word spotting, we provide a modified text box capable of generating an image from entered text. The user has the option of typing text, which is rendered in a font that mimics handwriting on the server side, or using their mouse to draw handwritten text directly in the box. Given a query image, the word spotting based approach will return the top N most similar cells in terms of handwritten contents, with the most similar being at the top. These results will not be perfect, with many of the returned images not matching the query text. However, using either of the methods used to enter text, we are able to acquire ASCII text from the user that we then associate with the results. If the user enters text in the box we can simply store this entered text. If the user uses the mouse to draw text, we convert the handwritten recognition problem to an online handwritten recognition problem where we know the path of the drawn strokes and are able to recognize the entered handwriting with much higher accuracies [2]. The acquired text is associated with the results clicked on by the user, assuming the users will tend to not click on incorrect results. In this way, the system passively transcribes the database, improving results, without the users being aware of it.

We have processed the state of North Carolina from the 1930 Census data using NCSA's Ember, a shared memory system with 4 systems containing 384 cores each and 2TB of memory. Containing approximately 60,000 scanned forms the segmentation and signature extraction took on average 132 seconds and 151 seconds per image respectively. The indexing of the signatures, organized into 2,280 separate indices for each reel and column, required on average 123 hours per reel. Requiring 98,000 CPU hours total, we estimate that to process the entire 1930 Census collection would require nearly 2.5 million CPU hours over a period of 288 days on this particular HPC resource.

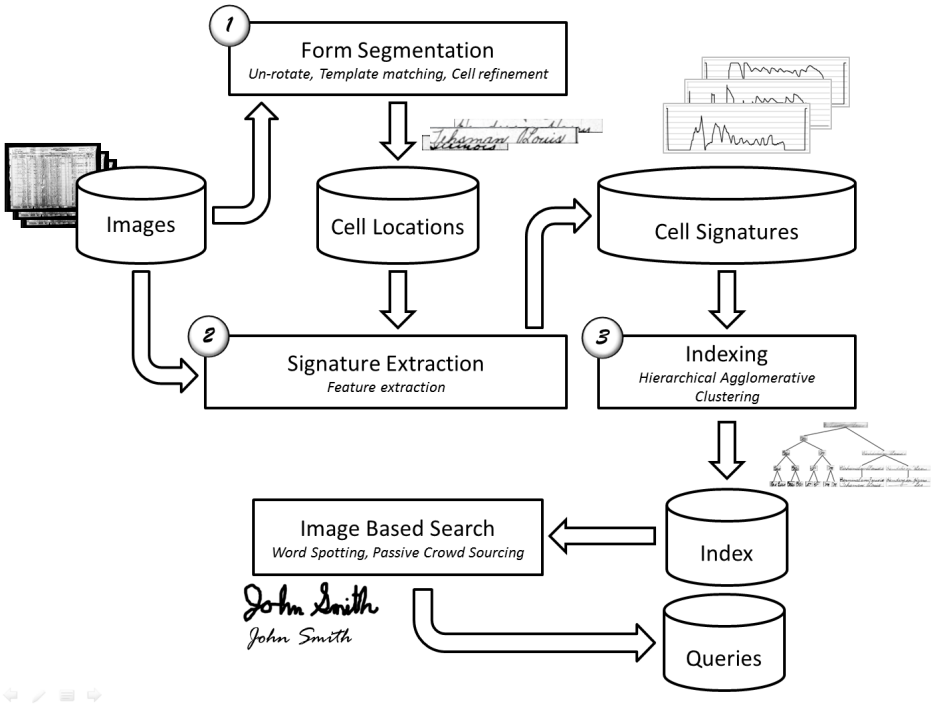

Figure 1. *A flow chart of the 3 pre-processing steps required to provide the image based search on the 1940 Census data. 1: The spreadsheet-like Census forms are segmented into individual cells. 2: A numerical signature is constructed to represent the handwritten contents of each cell. 3: A hierarchical index is constructed over the cell signatures.*

ACKNOWLEDGMENTS

This research has been funded through the National Science Foundation Cooperative Agreement NSF OCI 05-25308 and Cooperative Support Agreement NSF OCI 05-04064 by the National Archives and Records Administration (NARA). This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

REFERENCES

[1] T. Rath and R. Manmatha, "A search engine for historical manuscript images," *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.

[2] R. Plamondon and S. Srihari, "On-line and off-line handwriting recogntion: A comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.