10.1117/2.1201109.003833

Towards free and searchable historical census images

Kenton McHenry, Luigi Marini, Mayank Kejriwal, Rob Kooper, and Peter Bajcsy

Combining automation and crowd sourcing will provide access to archived handwritten forms.

Individuals and humanities researchers alike recognize the benefits of search services for censuses, which contain important information on ancestral populations.[1] In April 2012, the raw US census data from 1940 will be made available to the public for the first time in digital format. The census is being digitized by the National Archives and Records Administration and the US Census Bureau. Consisting of digitally scanned microfilm rolls, nearly 3.25 million photographs of the original census forms will be released (see Figure 1). The tasks of transcribing, organizing, and searching this very large >18TB corpus of images remains a resource-intensive task for other federal agencies. With databases of this type, a Soundex index, which encodes words based on how they sound to enable homophone matching, are often compiled. However, producing such an algorithm is a tedious and time-consuming process and will not be released with the 1940 data. On the day of the data release, various commercial entities will also begin transcribing the handwritten content of the images, a task that will take thousands of trained laborers anywhere between 6 and 12 months. As a result, access to the searchable, transcribed data will come at a cost to the public by these various companies. Here, we describe our approach to image-based information retrieval to avoid the costly transcription process.[2,3]

Our goal is to minimize the manual labor needed to transcribe handwritten entries in the census images and deliver a system capable of computationally scalable search services. Understanding the achievable accuracy and levels of automation depends on solving several problems related to scalability and data management. We endeavor to provide a completely automated search capability that can build more accurate transcriptions over time using passive and active crowd sourcing (see Figure 2). Commercial entities typically outsource manual transcription of census forms and host text-based search services

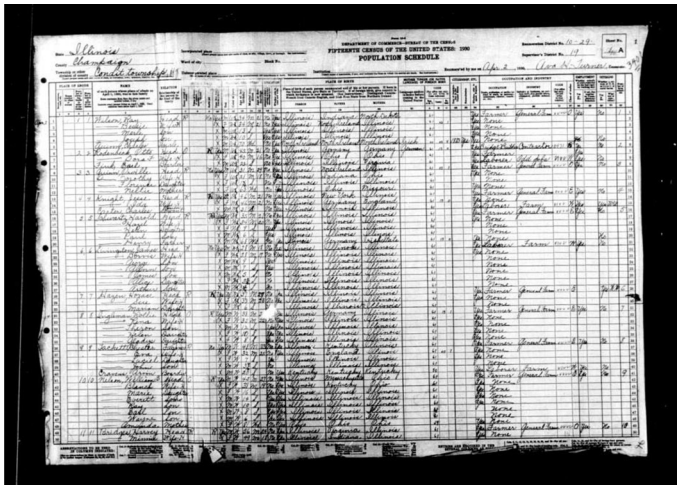

Figure 1. *A digitized census form from the 1930 US census.*

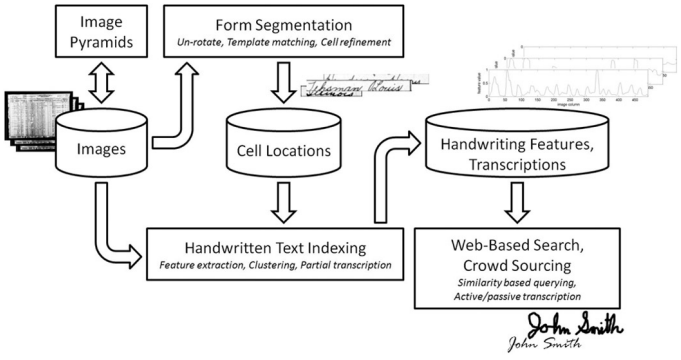

Figure 2. *The architecture of our proposed hybrid automated/crowd-sourcing system to provide access to content within scanned census forms. Image pyramids are used to pre-process the larger images to access small areas more efficiently.*

over those transcribed entries. In contrast, our approach uses form segmentation, handwritten text indexing, and web-based search and crowd sourcing to minimize the manual transcription of the images.

Continued on next page

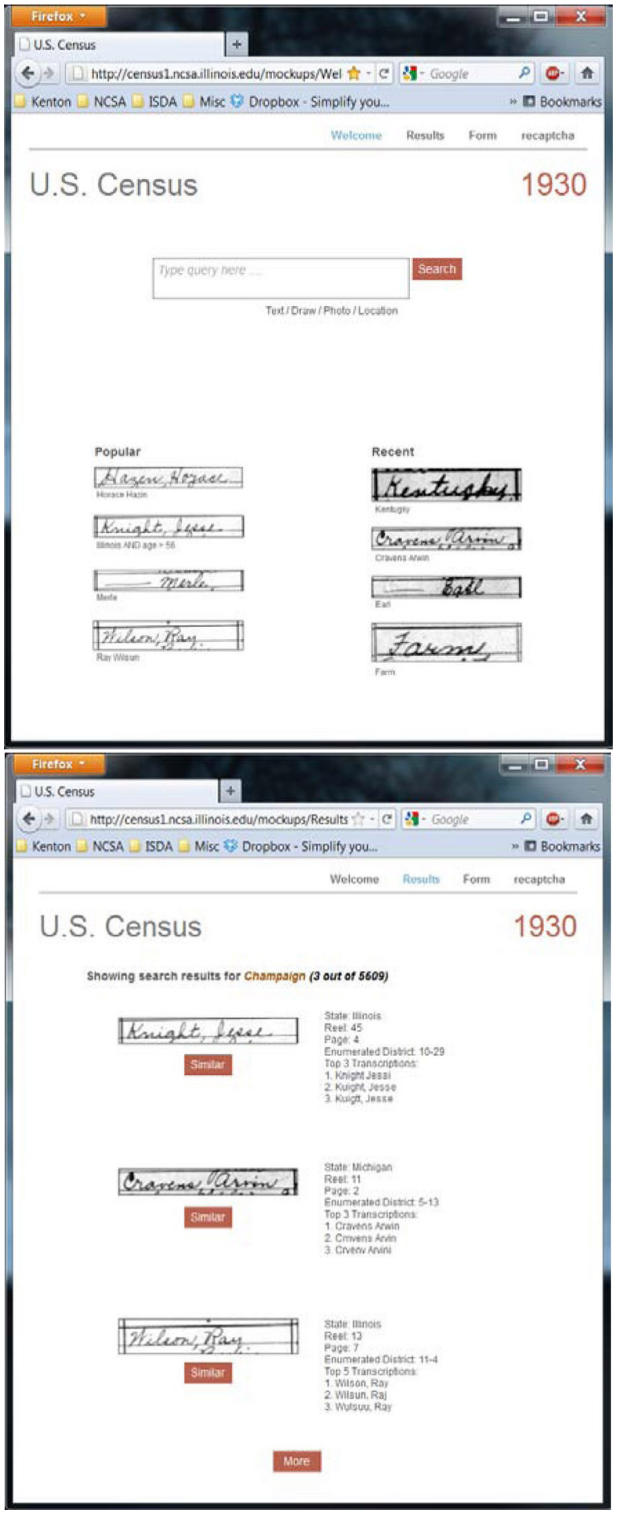

Figure 3. *(Top) The front end of our proposed system and (bottom) an example of what might be returned from a query.*

We are using the 1930 census images to develop our system. We began by segmenting the form to identify the grid cells. The scanned images often are rotated such that the form lines are not quite horizontal/vertical. Additionally, they can be smudged in places, torn at the edges, and have faint writing. We use a multi-step approach to overcome these problems. First, we reduce the size of the image, identify the long horizontal lines, and align the form horizontally/vertically. Next, we automatically find as many of the long horizontal and vertical lines as possible and superimpose a manually created template of the form. Finally, we refine each form cell position individually by essentially repeating this process of un-rotating and fitting. We found the forms to be fairly consistent within the 1930 census data, with most having a layout of either 25- or 50-row tables.

The next challenge was to recognize handwritten text, which is a difficult and active research problem. Rather than attempting to transcribe the data, we instead want a steerable means of retrieval. To do this, we are exploring a number of techniques based on the idea of word spotting.[4,5] This premise is based on the similarity of a set of features within two samples, one provided by users and one within the census images. We have evaluated the initial retrieval quality of the information obtained from the word spotting given a query image. Our preliminary results show that for fields with small lexicons (for example, 'yes,' 'no,' 'male,' 'female'), we can achieve accuracies up to 85.4%. For fields with larger lexicons—including places of birth—we have seen accuracies of 37.1% in terms of the top returned image matching the target and 76.1% in terms of an image in the top five matching the target. We are also exploring ideas[6] that extend word spotting beyond writing from the same person to include varieties of writers and even fonts that look like handwriting.

The final aspect of our approach relies on web-based search and crowd sourcing. Here, users will be provided with a modified text-box widget to enter text using the keyboard—in a font that looks like handwriting—or draw text using the mouse (see Figure 3). Features are then extracted from the entered text and used to find matches within the handwritten census forms. If a user spends a significant amount of time on a particular result from the search, we can associate typed text, or a transcription based on more accurate online handwriting recognition, as a possible transcription of the obtained result. Additionally, we can obtain transcriptions using something similar to reCAPTCHA,[7] where users are asked to type out contents from cells of the census forms before their own queries are executed. While CAPTCHA is typically used to distinguish

Continued on next page

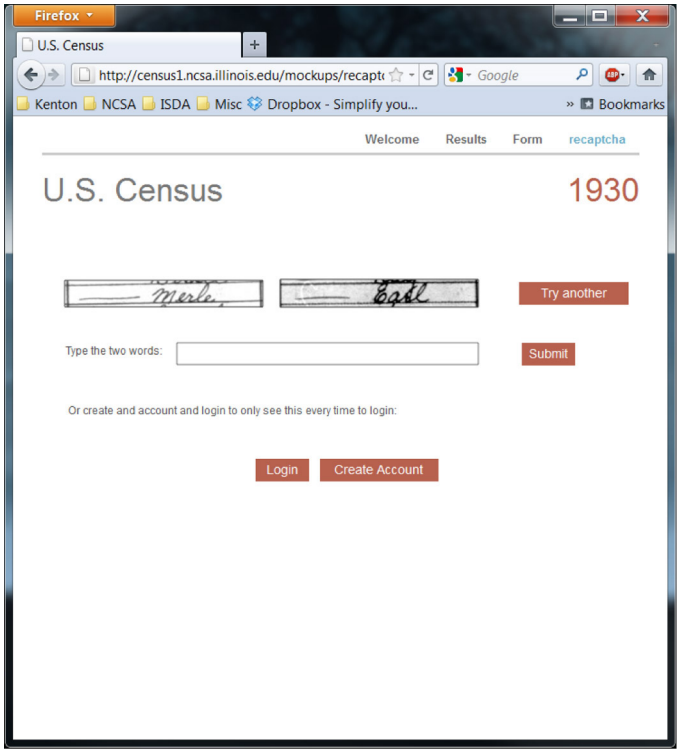

Figure 4. *A reCAPTCHA-like step presented to users before a query is carried out.*

machines from humans, a reCAPTCHA approach can gather transcriptions of the census forms (see Figure 4). This crowd-sourcing element will be investigated as a second phase, after our automation portions have been finalized.

In summary, our hybrid automation/crowd-sourcing approach aims to provide search capabilities over the image-based census data, potentially from the day the images are released. However, general difficulties in automating handwriting recognition will limit its accuracy. Incorporation of passive and active crowd-sourcing elements will improve the accuracy of our systems over time. We are currently working on a number of challenges, including further pre-processing of form cells to remove noise. Our next important stage will be to build an index of the ∼7 billion form cells, which is crucial for efficient access. However, of the word-spotting techniques we tested, the best results use a non-linear comparison that does not lend itself to indexing. We are currently investigating alternative methods that are indexable, as well as using high-performance computing resources to perform a one-time, large pre-processing step to hierarchically cluster the data (requiring 4.9×10^{19} comparisons). Finally, we will investigate how best to associate the passively crowd-sourced transcriptions with the results based on user behavior.

Author Information

Kenton McHenry, Luigi Marini, Mayank Kejriwal, and Rob Kooper
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
Urbana, IL

Kenton McHenry received his PhD in computer science from the University of Illinois at Urbana-Champaign in 2008 after completing his BS from California State University, San Bernardino. He is currently a research scientist.

Peter Bajcsy
National Institute of Standards and Technology
Gaithersburg, MD

References

1. M. Anderson, *The census, audiences, and publics*, **Soc. Sci. Hist. 32**, pp. 1–18, 2008. doi:10.1215/01455532-2007-011
2. K. McHenry and L. Marini, *Searching the 1940 census*, **75th Ann. Mtg. Soc. Am. Arch.**, 2011.
3. K. McHenry and L. Marini, *Towards free and searchable access of the 1940 census data*, **Nat'l Assoc. Gov. Arch. Records Admin. Council State Arch. Joint Ann. Mtg.**, 2011.
4. R. Manmatha, C. Han, and E. M. Riseman, *Word spotting: a new approach to indexing handwriting*, **Proc. Comput. Vision Pattern Recognit.**, pp. 631–637, 1996. doi:10.1109/CVPR.1996.517139
5. T. M. Rath and R. Manmatha, *Word image matching using dynamic time warping*, **Proc. Comput. Vision Pattern Recognit. 2**, p. 521, 2003. doi:10.1109/CVPR.2003.1211511
6. J. A. Rodriguez-Serrano, F. Perronnin, J. Llados, and G. Sanchez, *A similarity measure between vector sequences with application to handwritten word image retrieval*, **Proc. Comput. Vision Pattern Recognit.**, pp. 1722–1729, 2009. doi:10.1109/CVPR.2009.5206783
7. L. von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, *reCAPTCHA: human-based character recognition via web security measures*, **Science 321**, pp. 1465–1468, 2008. doi:10.1126/science.1160379