Census

In two words the purpose of this project is providing “searchable access” to digital archives. In particular we are interested in the many digital archives of handwritten forms that currently possess no viable or practical means of searchable access. Today access can be provided to archived handwritten forms by digitally scanning them and making them available on the web as a collection of digital images. This type of access is only of limited use however as currently in this form users have no easy way of searching through the individual forms other than by going through them one by one manually. Given that these collections are almost always large, in the terabytes of data made up of millions of images, the process of searching through the data manually is for all practical purposes impossible for an individual.

The motivation for our work is the decadal release of US Census forms coming up in April of 2012 for the 1940 Census. The digital scanning of 1940 US Census microfilms by the US Census Bureau and the National Archives and Records Administration (NARA) has been estimated to result in 3.25 million images and approximately 125 terabytes of raw image data. Compressed copies of the images, approximately 18 terabytes worth, will be made available over the internet for the first time as the primary source of the Census information.

Upon the release of the census information various companies and organizations will begin the long tedious task of manually transcribing the contents of the handwritten forms. As transcribed digital text the Census information will become searchable. To do this the companies/organizations will employ anywhere from thousands to hundreds of thousands of people to manually look at each word within the forms and retype them as digital text. The process will take between 6 to 9 months to complete. Due to the large investment in human labor, these companies will provide the searchable content for a fee. At the moment there is no low cost, near free, alternative to human transcription for providing searchable access to information within digital archives of handwritten information. It is here were our work focuses.

We aim to provide a means of providing low cost searchable access to digital archives where there is currently nothing. The field of Computer Vision, which deals with the extraction of information from images, is by no means at the state required to provide perfect automated transcription of the contents of these forms. However, our goal is not perfect machine driven transcription, but instead providing some form of searchable image based access where no searchable access would exist otherwise.

Towards this goal of providing low cost searchable access to digital archives of handwritten forms we investigate and develop a hybrid automated and crowdsourcing approach. The automated portion, utilizing a technique known as Word Spotting, will provide immediate searchable image based access to the content within digital handwritten forms. The crowdsourcing component, made up of both active and passive crowdsourcing elements, will accumulate traditional transcriptions over time. Of particular interest is the passive crowdsourcing element which would acquire these transcriptions without the user being made aware of the fact that they are carrying out this job. As more and more users use the system and transcriptions are acquired over time the system will gradually shift from solely image based search to a combination of image based search and text based search. Given the limitations of the current state of computer vision our approach emphasizes the human in the loop. Even within the automated portion we provide novel interfaces so that the human can search in a manner that is more readily tractable by the computer. In addition to the computer vision and crowdsourcing elements we must deal with the scale of the problem at hand, that is millions of images and terabytes of data. Towards these ends we also investigate the use of scalable distributed databases, efficient means of providing high resolution images over the web, and means of indexing the large amount of text within an archive of digital images.

Recently Updated

Census
Mar 04, 2014 • updated by Kenton McHenry • view change
Census
Apr 15, 2011 • updated by Rob Kooper • view change
Census
Apr 15, 2011 • created by Rob Kooper