

Image Registration and Text Recognition for Structured Census Documents

Krishna Subramanian, Huaigu Cao, Xujun Peng, Rohit Prasad, Prem Natarajan
Raytheon BBN Technologies,
10 Moulton St.
Cambridge, MA 02138

ABSTRACT

In this paper, we present our work on developing a system for registration and recognition of structured census documents. Information extraction from these documents present many challenges, for instance, table registration, cell extraction, binarization, and recognition of handwritten text. This paper mainly deals with table registration. It details the approach and algorithms we developed for unsupervised registration of tables given a set of templates. The algorithm is also capable of detecting the presence of a template in a page before proceeding to register it. No restrictions are placed on the position or the size of the table in a page in comparison to those of the template and are robust to skew and minor amounts of non-linear distortions in the scanned page. We then proceed to outline our overall system for information extraction from tabular pages using the BBN Byblos Optical Handwriting Recognition (OHR) system. We present preliminary results for table registration using our approach.

Categories and Subject Descriptors

Optical handwriting recognition, Table registration, Information extraction, document analysis

Keywords

Table registration, Optical handwriting recognition

1. INTRODUCTION

Many historical documents such as census records have handwritten text recorded in structured tables. Manual extraction of information in these tables into a computer database is an expensive and labor intensive task. With recent advances in document imaging technologies, including optical character recognition of handwritten text, it has become possible to improve the accuracy and efficiency in extracting information from such structured documents. In this paper, we describe an initial prototype that we have built for recognizing the content in scanned census records.

Tabular census records present several challenges to the state-of-the-art in image registration and optical character recognition. These challenges go well beyond the generic challenges in recognizing handwritten text such as variation in writing styles, writing instruments, slant, character segmentation, etc. Poor scanning, stroke bleeds, distorted cell boundaries, folds, etc. make the registration task of finding the locations of the cells and assigning text to a particular cell extremely difficult. The rest of this paper is organized as follows: In Section 2, we describe the corpus used by our system and the challenges for information extraction from these documents in Section 3. This is followed by a detailed description of our algorithm for performing table

registration on these documents in Section 4. Along with the description, we also provide many figures illustrating our approach. A brief overview of the rest of the information extraction pipeline including a brief description of the BBN Byblos OHR system [1] follows in Section 5. We show some preliminary experimental results in Section 6 and our conclusions in Section 7.

2. CORPUS DESCRIPTION

In this paper, we are interested in extracting information from structured documents, in particular, scanned tabular documents from the US and Mexican census records. The scanning was done at 300 dpi. The United States Census records were from the year 1930 and span 35 states while the Mexican records were also from the year 1930 and span 15 states. Each of these documents has a set of tables whose cells have handwritten text. The rows and columns in these tables are demarcated by horizontal and vertical lines. The tables in each document come from a set of known templates. For each document, a subset of cells was transcribed. The final goal for information extraction was to reproduce the manual transcriptions for this same subset of cells automatically. The transcribed cells could be used for training and development of the OCR engine. Apart from these annotations, we also used downloaded content from the internet to building our names database in turn used for building our OCR dictionary and language model.

3. Challenges for Information Extraction

There are quite a bit of challenges pertaining to the scanning process: (a) pages have different skew, (b) the scanned size of the tables is inconsistent, (c) some documents were clean while others had significant amounts of salt and pepper noise added while scanning, and (d) the intensity of rows and columns varies significantly. The original census records are stored folded by half around the center from top to bottom. During scanning, the page was unfolded and flattened against the scanner. These folds are clearly visible on the scanned document image. Due to inconsistent flattening during scanning, each document has variable amounts of non-linear shear especially in regions where the original document is folded. This non-linear shearing complicates the registration process and precludes using certain algorithms like correlation based techniques for template matching. The other challenges pertaining to cell extraction and OCR are: (a) the writing instrument used to record the census and the pressure exerted while writing resulted in strokes with large variation in thickness and intensity complicating the binarization process, (b) the handwriting styles varied greatly and presented challenges not just to the OCR engine but also to the cell extraction process because, depending on the handwriting style, strokes from text in one cell flowed into adjacent cells.

4. Technical Approach for Table Registration

The block diagram for table registration is shown in Figure 1. The registration process can be conceptually divided into six stages: (a) Page cleanup removes large connected components from the scanned document, (b) The cleaned up document divided into vertical blocks spanning the entire height of the image, (c) Each block is individually analyzed for salient points such as the four corners and cross, i.e. intersection of row and column lines, (d) The detected salient points from each block is composed into the

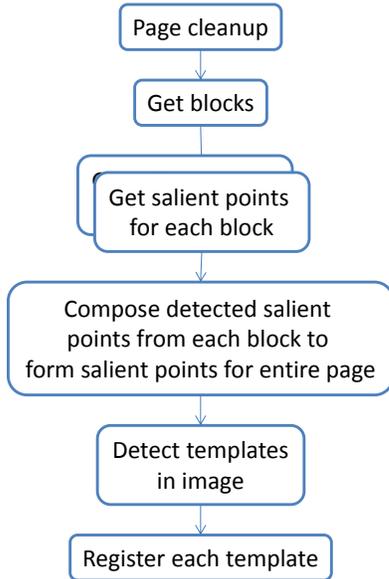


Figure 1: Block diagram for Table Registration

original image coordinate space, (e) Each template is detected from the salient points and loose constraints on its location, (f) Each detected template is then registered. We now proceed to describe some of these stages in more detail.

4.1 Page cleanup

The first step is to clean the image to get rid of large components appearing at the four boundaries of the scanned image. For cleanup we do the following: (a) significantly downsample the image so that all the text is consumed, (b) Perform morphological OPEN operation using a rectangular mask of suitable size so that only the large connected components remains. We then expand the remaining large connected components slightly in each direction. This expanded connected component image is then subtracted from the original image so the resulting image only contains components of interest.

4.2 Detect salient points

We assume that the block for which salient points need to be detected is has little non-linear distortions and can be treated as an affine transform of the original table for all intents and purposes.

The block diagram for detecting salient points is shown in Figure 2. We first detect columns and then correct for skew resulting in the columns being very nearly vertical. We detect 5 types of salient points in the image, namely, the four corners and crosses (intersections of rows and columns). Unfortunately, popular approaches for corner detection, including the FAST detector, Harris-Stephens and Shi-Tomasi algorithms all break on such

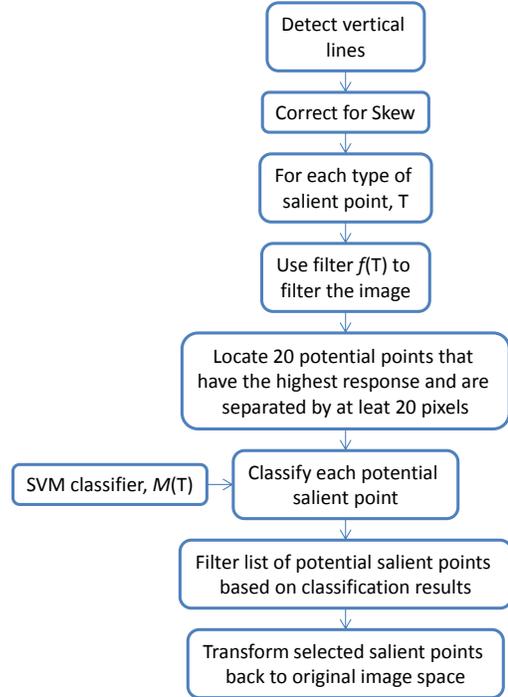


Figure 2: Block diagram for detecting salient points

images. Instead, we have developed a novel learning-based approach to detect corners and crosses. For each type of salient point T , we designed a specialized template that responds well to these locations. The five templates are shown in Figure 3. The positive and negative valued pixels are shown in yellow and green respectively. Zero valued pixels are shown by blue. Very negative pixels are shown in red. The origin for each template is the point of intersection of the horizontal and vertical yellow lines. The size of the template is 20×20 pixels and the thickness of the green strip is 3 pixels and the thickness of the yellow strip is 1 pixel. This template is used to filter the image. The top 20 points with highest responses and also separated from each other by at least 20 pixels is chosen. These points are then further filtered by using the SVM classifier that is trained using the 20×20 pixel region around each pixel centered at the intersection of the yellow lines in Figure 3. The SVM-filtered salient points are then transformed back to the original image coordinates. The detected salient points for a scanned document created by transforming the detected salient points for each block to the image coordinate space is shown in Figure 4.

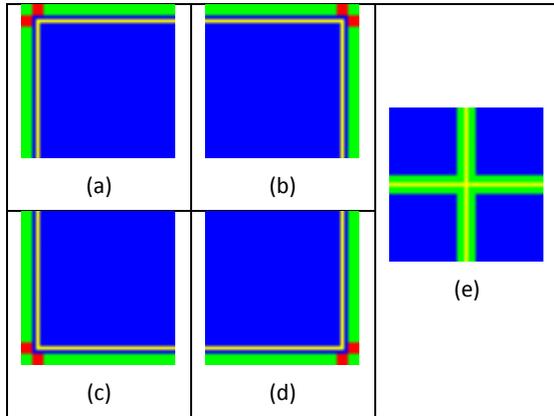


Figure 3: Templates used for detecting salient points. (a) Top left, (b) Top right, (c) Bottom left, (d) Bottom right, (e) Cross. The values for each color are: Red=-10, Green=-1, Yellow=1, Blue=0

4.3 Detect templates in image

The salient points (corners and crosses) for the template are assumed to be known. The approximate location of the template

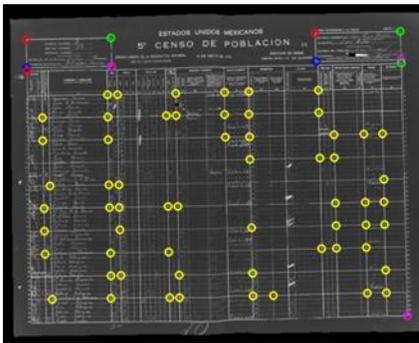


Figure 4: Detected salient points

in scanned document is also assumed to be known. We first rotate the entire document image and associated salient points based on the median of detected skew angles for each block. The salient points for the document are then used to create an impulse response image. This impulse response image is convolved with a Gaussian cloud. This image, for the same document in Figure 4, is shown in Figure 5. The Gaussian cloud image is used to approximately localize the template in the document filtering the sampling the Gaussian cloud image with and impulse response image constructed from the salient points in the template and looking for the point with the highest response in the neighborhood that is known to contain the template. The partial registration result for one of the templates is shown in Figure 6. We then perform a constrained nearest neighborhood search to find pairs of salient points in the template and the image such that

these points are of the same type, i.e., the right corner matches with the right corner. Also, we ensure one point is within a 30



Figure 5: Gaussian cloud created from the detected salient points

pixel radius of the other. At this stage, if the number of pairs found is less than 2, we give up searching for the template in the image. Once the salient points are aligned, we estimate a rigid

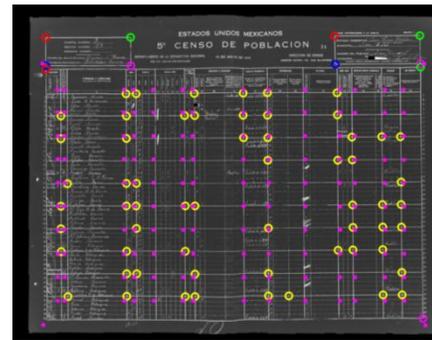


Figure 6: Localized and partially registered template

affine transform between these pairs of points. After transforming the detected salient points in the document, we repeat the nearest search again and improve our transform estimate. At this point, the images are very close the registered. We refine our registration to account for the non-linearity in the scanning process. Each point in the horizontal axis is mapped to another point along the horizontal axis by estimating a b-spline interpolation model using the pairs of salient points. The final registered and extracted tables for the document in Figure 4 are shown in Figures 7a, 7b and 7c. Once the template is registered, since we know all the coordinates of the cells we want to extract in the template, it is easy to extract these coordinates from the registered document image. An example of an extracted cell is shown in 7d. The extracted cell is then cleaned for border lines. In the current prototype, we do not account for noisy strokes from neighboring cells and treat each cell as an independent unit.

