Automatic Extraction of Phrase-Level Map Labels from Historical Maps

Haowen Lin, (Mentor: Yao-Yi Chiang)
University of Southern California
haowenli@usc.edu

1. INTRODUCTION

Historical maps are important resources for various kinds of studies, providing insights for natural science and social science studies such as biology, landscape changes, and history [1]. However, text recognition on maps remains a challenging task because map usually has a complex background in which textual content appears in numerous colors, fonts, sizes, and orientations. Even if we were able to acquire perfectly recognized words and characters automatically, it is still difficult to generate useful information because individual words are not meaningful. For example, a typical result from OCR scanning or manual map digitization is that each recognized bounding box only contains a single word (Figure 1).

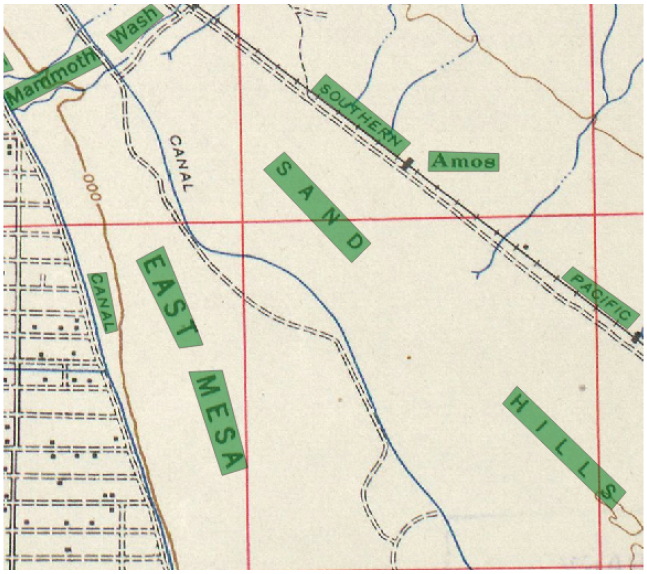

Figure 1: Example of recognized bounding boxes (green polygons)

Bounding boxes of the same phrases could be far away from each other, increasing the difficulty of linking them (e.g., SAND and HILLS, SOUTHERN and PACIFIC in Figure 1). This paper presents an automatic approach that combines single words extracted from historical maps into meaningful phrases, which represent complete location descriptions and can be used to link historical sites to other datasets. Our algorithm first combines textual and spatial features of individual map words to evaluate the potentiality of connecting two words. Then the algorithm trains a support vector machine to adjust the weight of each feature. This algorithm is potential to improve digital map processing by increasing the automation of text extraction on maps.

2. APPROACH

Table 1: Input Data and Output Data for Polygons in Figure 1

Input Data (Geo polygon)	Output Data
Mammoth	Linking with "Wash"
Wash	Linking with "Mammoth"
EAST	Linking with "MESA"
MESA	Linking with "EAST"
SAND	Linking with "HILLS"
HILLS	Linking with "SAND"
SOUTHERN	Linking with "PACIFIC"
Amos	No linkage
PACIFIC	Linking with "SOUTHERN"

The input data are the minimum bounding boxes for each word on maps. The output data is whether there exists a link for a pair of bounding boxes to constitute a phase. We assume all textual contents of the input data are perfectly transcribed. Table 1 presents the input data and ideal output data for bounding boxes in Figure 1.

2.1 Generating Feature Abstraction

Our algorithm uses four heuristic features to determine if two words should be linked to constitute a phrase. The features include boundary distances between two polygons, the text area for each character inside the bounding box, capitalization of the word and text contents.

2.1.1 Boundary Distance

Under most circumstances, bounding boxes with words in the same phrases are located nearby. Therefore, relative distances between two polygons can be a significant indicator for measuring word connection. We compute the distance between every line segment pairs on the boundary of every two bounding boxes and record the shortest one as the boundary distance. We use boundary distance instead of center-to-center distance because the polygons themselves could occupy a wide area and increase calculation errors. Boundary distances do not necessarily define whether the selected bounding boxes are in the same phrase or not, though.

2.1.2 Text Area for Each Character

Each map data consists of a varying number of text fonts. Words in the same phrases, even though separated, do not change their text fonts. However, identifying text font from maps with complicated layouts are challenging and time-consuming. Historical maps usually contain handwritten text also increase the difficulties for map label recognition [2]. To simplify the process and reduce errors, we use the area of each bounding boxes divided by the number of characters to distinguish text fonts.

2.1.3 Capitalization

There are three situations for case-sensitive textual contents on the map: 1) All letters are uppercase, 2) All letters are lowercase, and 3) Words are combinations of uppercase and lower letters. Having the same capitalization is the prerequisite for connecting two polygons. For example, "SAND" and "HILLS" in Figure 1 are both capitalized words because they are in the same phrases while "Salton" and "sea" will not be linked together because they have different capitalizations.

2.1.4 Textual Content

Textual contents are useful data sources to improve the accuracy of word linking. For example, in the United States Geological Survey (USGS) Historical Topography Maps, if the connected words match any of the place names from the USGS gazetteer, we mark the connection as a "confident linking." Additionally, bounding boxes with exactly same text content such as "mountain" should not appear in the same phrases.

2.2 Applying Training Algorithm

Finding the "correct" threshold for each of the features in Section 2.1 to connect single words would not be reliable, and this problem has an intuitive implication to use Support Vector Machines (SVMs) in the classification settings. SVMs are supervised learning models that are useful in linear classification.

3. PRELIMINARY RESULT

We evaluated our algorithms on real-world data from two historical map resources: Ordnance Survey and USGS Historical Topography Map. We worked with two sets of bounding boxes taken from these databases:

Set 1: 205 bounding boxes manually transcribed from USGS maps

Set 2: 758 bounding boxes manually transcribed from USGS and Ordnance Survey maps

For each set of bounding boxes, we manually digitized the maps and created ground truth data, storing the text of words and related phrases with polygons. For the current research, we only used USGS National Geographical Names to generate the input gazetteers.

We used 4 parameters to train the SVM model: boundary distances (float numbers), text area for each character (float numbers), Capitalization, (Boolean values, true if the two polygons have the same capitalization), text content (Boolean values, true if the two polygons have same text content). We used Set 1 for training and Set 2 for testing.

We used Precisions and Recall to evaluate the performance. If two words representing same phrases were connected, we labeled this association as "correct linking," otherwise marked it as "wrong linking." We assume there would be a linkage between every pair of words in the ground-truth phrases. For example, in phrases "Old Cruikshank Ranch," the algorithm added three connections: "Old" and "Cruikshank," "Old" and "Ranch," "Cruikshank" and "Ranch" into total linkages for calculating the precision and recalls. Table 2 presents the experiment results.

Table 2: Experiment results

Result	USGS 60 Inches-Salton	Ordnance Survey 60 Inches	USGS 15 Inches-Brawley
Precision	91.56%	91.67%	79.31%
Recall	40.42%	38.60%	32.85%
Total Phrases	95	84	134

From Table 2, we can see that the algorithm showed excellent performance on the precision with accuracy over 79% for all types of maps. The mistake usually occurred when multiple polygons with similar text font but representing different phrases are aggregated or overlapping with each other (Figures 2 and 3).

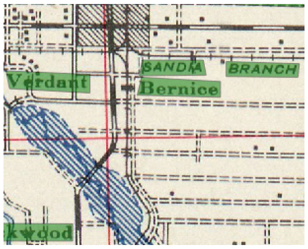

Figure 2: Aggregated polygons

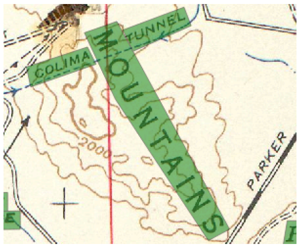

Figure 3: Overlapping polygons

The low recall showed that the algorithm missed out many linkages. One reason is that some bounding boxes with words in same phrases are in a great distance ("EAST SIDE HIGHLINE CANAL" in Figure 4). In this case, roads, rivers, transmission lines are critical indicators of linking, which were not used in the algorithm. Another reason is that bounding boxes also do not remain a fixed orientation, thus increases the challenge for linking ("San Felipe Creek" in Figure 5).

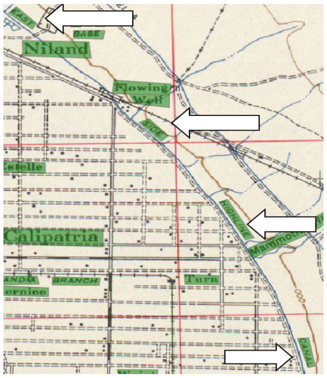

Figure 4: Example of bounding boxes in a great distance

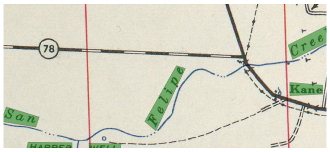

Figure 5: Example of curved bounding boxes

4. DISCUSSION AND FUTURE WORK

We presented an algorithm that combines textual and spatial information of map words to automatically generate meaningful place information. Some directions for future work including generating more features for evaluation and trying other types of machine learning algorithms to deal with the situation when the map label is curved. We also plan to adaptively link the words by removing connected bounding boxes from the map and then applying the algorithm to link the rest in order to improve the recall.

5. ACKNOWLEDGEMENT

This research is supported in part by the National Endowment for the Humanities (Grant No.: NEH PR-253386-17) and in part by the USC Undergraduate Research Associates Program.

6. REFERENCES

[1] YY. Chiang. Unlocking Textual Content from Historical Maps-Potentials and Applications, Trends, and Outlooks. International Conference on Recent Trends in Image Processing and Pattern Recognition, pages 111-124. Springer, 2016.

[2] M. Heliński, K. Miłosz, and P. Tomasz. Report on the comparison of Tesseract and ABBYY FineReader OCR engines. 2012.