2019 Harvey Mudd College Mathematics Clinic

Detecting Environmental Hazards in Historical Maps

Introduction

This project analyzes historical map data, which is useful for showing temporal changes of buildings and businesses. Our focus is to highlight and interpret important words on the maps indicating environmental hazards.

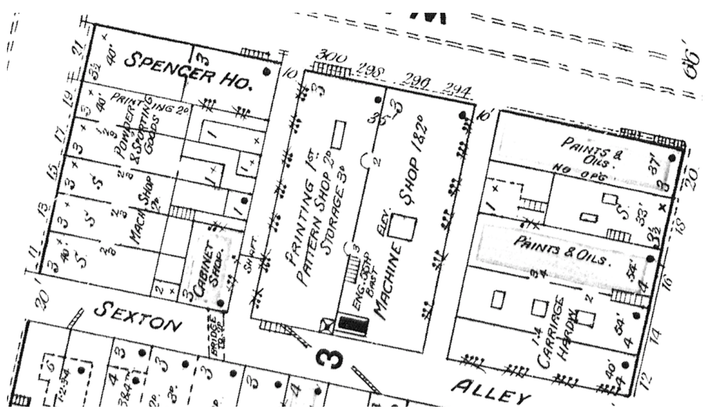

Figure 1: A section of a historical map.

Methods

Image Processing

Figure 2: A section of a binarized map before (left) and after (right) line removal.

This step consists of four main components.

1. Convert the image to binary to make it easier to process.
2. Apply an algorithm developed by last year's clinic team to remove long lines corresponding to street curbs and parcel boundaries (Figure 2).
3. Identify pixel islands, or fully connected groups of pixels.
4. Filter them through a junk classifier to find alphanumeric symbols (Figure 3).

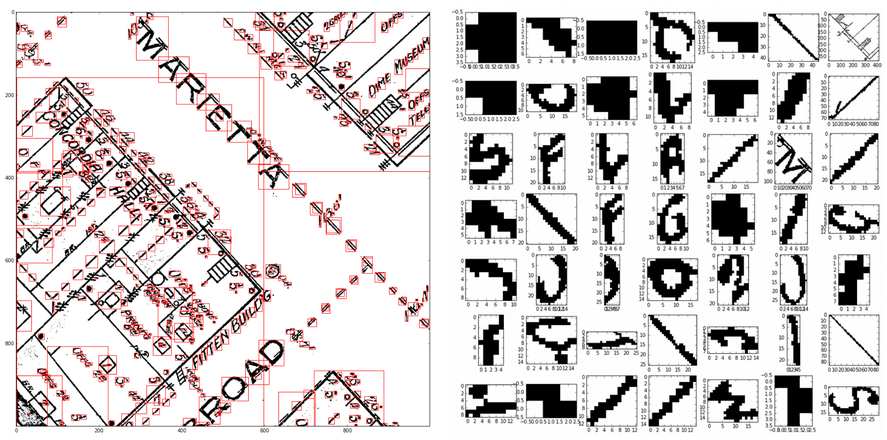

Figure 3: A section of a map with bounding boxes around the pixel islands (left) and an array of isolated pixel islands (right).

Optical Character Recognition

Using a set of 20,000 hand-labeled pixel islands and a set of 100,000 labeled EMNIST digits, we trained a rotation-invariant convolutional neural network model to identify symbols. An example output is shown in Figure 4.

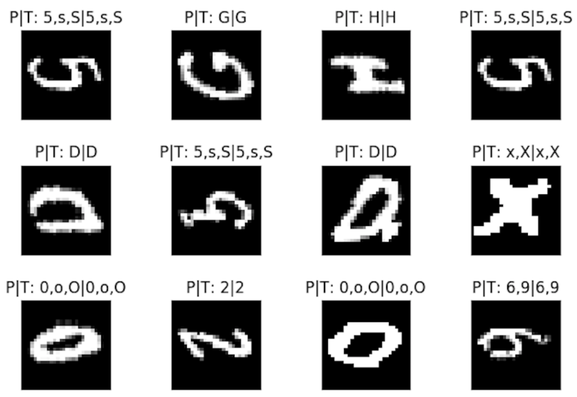

Figure 4: Sample output of OCR, where P stands for the predicted class and T stands for the true class.

Word Finding

After identifying individual pixel islands, we use a KD-tree based algorithm to group symbols together. We then identify the words most likely to correspond to a given group of symbols using a custom spell-checking model.

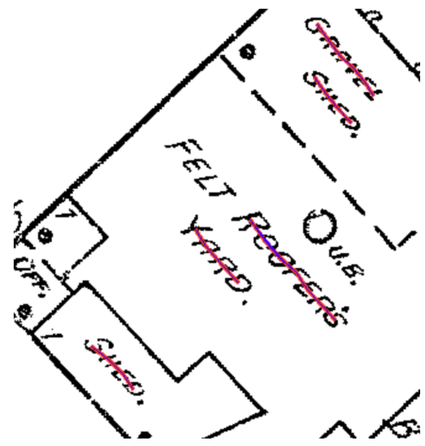

Figure 5: A section of a map with symbols corresponding to the same word grouped.

Conclusion

We successfully developed a solution that can take a scanned raster map and identify potentially important words. The ability to convert a scanned digital map to an easily manipulable and searchable format will allow EDR's clients to more easily identify potential environmental issues on the historical maps.

Acknowledgments

We would like to thank our liaisons Zachary Fisk, Paul Schiffer and Richard White for their valuable guidance throughout the process and our advisor, Professor Nicholas Pippenger, for supporting us during this project. We are also grateful to Clinic Director Professor Weiqing Gu and Clinic Coordinator DruAnn Thomas. We would also like to thank the people who helped label training data for our OCR model.

Team Members

- Nathaniel Diamant
- Mackenzie Kong-Sivert
- Jacky Lee (Spring PM)
- Vivaswat Ojha
- Kinjal Shah (Fall PM)
- **Faculty Advisor** Nicholas Pippenger
- **Liaison** Zach Fisk
- **Liaison** Paul Schiffer
- **Liaison** Richard White