GOOD SYSTEMS
A UT Grand Challenge

Machine Learning for GIS (ML4GIS):
Striving for Scalable Processing of Scanned Map Images

TEXAS Libraries
The University of Texas at Austin
University of Texas Libraries

The University of Texas at Austin
School of Information

Michael Shensky[1,4], Danna Gurari[2,3], Aaron Choate[1]
[1]UT Libraries, University of Texas at Austin [2]Dept. of Computer Science, University of Colorado Boulder [3]School of Information, University of Texas at Austin [4]Department of Geography, University of Texas at Austin

Introduction

The UT Libraries manages a growing collection of over 60,000 scanned map images that need expanded metadata and georeferencing to support discovery and use. It takes ~15 minutes to manually georeference and develop metadata for one map. It would take almost 8 years for a single person working 40-hour weeks to process all 60,000+ maps. Other organizations, institutions, and local government agencies, like the City of Austin, likewise maintain large map collections and would benefit from a automated method of extracting useful information from their maps.

Objective

The goal of the ML4GIS project is to leverage crowdsourcing to efficiently produce map annotations that can be processed to generate metadata, georeference maps, and ultimately be used to train a machine learning model that can be used as part of a scripted process for generating annotation data for maps without manual map interpretation being required.

Work Completed

We have developed the ML4GIS open source annotation tool using NodeJS and ReactJS to facilitate annotation of scanned map images (Zhan & Chao 2021). This tool is customizable so that it can be easily modified to accommodate specific annotation tasks on different map collections. It has been designed to be deployed via a crowdsourcing platform like Amazon Mechanical Turk so that it is possible to scale up the process of generating a large annotation dataset that can be used to train a machine learning model for identifying specific information in scanned map images.

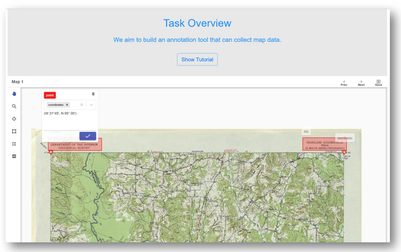

Figure 1. Screenshot of ML4GIS annotation application

Current Work

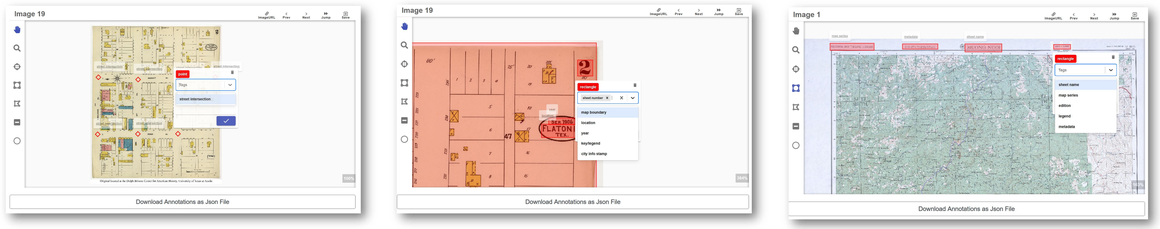

Figure 2. These screenshots show the ML4GIS application customized for annotation of Sanborn maps (left and center) and Army Map Service maps (right)

We have started using the ML4GIS annotation tool to create annotation data for scanned map images and have verified that the annotations can successfully be exported to JSON format and used to georeference and develop metadata for scanned map images using a scripted process. As seen in figure 2 above, we have recently developed different configuration settings for the ML4GIS annotation application so that it can be streamlined to improve the efficiency with which particular map elements can be annotated for different map collections. We plan to develop additional configuration settings for specific map elements and other map collections so that we can make the annotation process as quick and simple as possible for the crowdsourced workers we soon intend to pay on a per map basis to annotate maps from the UT Libraries' collections.

Because the viability of a computer vision model is dependent on the quality of the data used to train the model, we are next preparing to conduct a quality assessment of annotations generated by crowdsourced workers. This will be conducted by comparing the annotations for identical map images that have been separately worked on by UTL staff and crowdsourced workers on Amazon Mechanical Turk. Members of the research team are now going through the process of creating highly accurate annotations for a set of hundreds of maps randomly selected from the UT Libraries' map collections. These annotations will be treated as the groundtruth for determining the accuracy of crowdsourced annotation data.

amazon mechanical turk

Figure 3. Annotation JSON data generated with the ML4GIS application

Next Steps

If we are able to determine that crowdsourced annotation data generated by workers on a per map basis is of sufficient quality for training a computer vision model, we plan to scale up the crowdsourced generation of annotation data for the UT Libraries' scanned map images. We will continue using this approach to annotation creation until we have a sufficiently large annotation dataset to successfully train a machine learning vision model to identify specific types of information found on map images using computer vision. Once we have a viable model, we will apply it to the task of identifying information in scanned map images that have never been manually processed or annotated. Using a computer vision approach to identifying obtaining valuable information about map names, map labels, map extents, and key features found in maps, we can then use approaches like OCR (optical character recognition) and scripted processing of text to generate metadata and the coordinate information necessary for georeferencing of scanned map images.

Expected Output

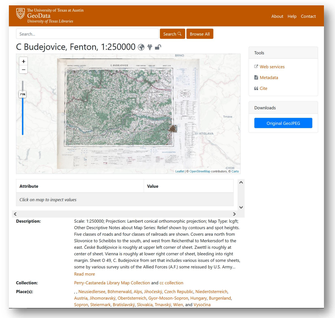

Figure 4. A georeferenced map with metadata from the UT Libraries' PCL Map Collection on the Texas GeoData portal

Our desired output is georeferenced scanned map images with separate metadata files describing the content of each map. It is our goal to that these processed and enhanced maps match the quality of maps that have been manually processed and which are currently shared through the UT Libraries Texas GeoData portal (see map pictured above). The success of this project would allow for us to use a machine learning approach to efficiently detect significant information on scanned map images so that these maps can be georeferenced and described much more quickly than via a manual process. This will allow the UT Libraries to dramatically increase the pace at which new maps are added to the Texas GeoData portal (https://geodata.lib.utexas.edu). If this is successful, we will be able to share the annotation application, code, and methodology with other institutions and organizations who can benefit from using these resources on their own map collections.

Acknowledgments

We would like to acknowledge all graduate student assistants who have contributed to this project: Alyssa Zhan, Patrick Chao, and Supreeya Karn. We would also like to thank our project partners Ross Clark at the City of Austin and Jennifer Hecker and Mike Miller from the Austin History Center who contributed maps to this project. Finally, we would like to acknowledge the Good Systems funding that has supported this project.

References

Zhan, A. and Chao, P. 2021. ML4GIS. GitHub Repository. Accessed at https://github.com/alyssazhan/ML4GIS

The University of Texas at Austin
Bridging Barriers
Office of the Vice President for Research