IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Project 3A - Digital automation of marine data

Auteurs Xue WANG, Runlu QU, Zheng YUAN, Hang ZHANG

Supervisor: Cécile BOTHOREL

Clients: Thierry Schmitt, Julian Le Deunff, Eric Le Guen

Key words : : Numerisation, Automate recognition, Image processing, Deep Learning

Executive Summary

1. Context of the project

This project is proposed by Shom and implemented by our team, a group of students from IMT-Atlantique. Shom, the first official hydrographic service in the world (1720), is an administrative public administration under the supervision of the Ministry of the Armies. Shom has a large number of bathymetric maps on which are handwritten bathymetry annotations. In order to exploit these documents, digitization of these maps is necessary. The goal of this project is to recognize automatically all the figures(numbers of the bathymetry) on these maps.

2. Problem analysis of the project

The main difficulty of this project is that there are a lot of noises around targets and writing styles on maps of different ages are quite different and we don't have a training set for all these writing styles.

3. Research and implementation of algorithms

We first worked on map denoising, we tried the 8-neighborhood method to remove dot-like noises and it works great; for line-like noises such as contour line, we tried HoughLines and Fast Line Detector, finally we found Fast Line Detector works well. But both algorithms may remove parts of targets, so we made a pre-positioning processing by copying suitable target areas to a new image and adding a black circle on each target, then we used FindContours function to locate all targets, with all these steps we can successfully locate 60%-90% targets on maps.

Once we get targets, the next step is to recognize them. We tried 2 ideas, the first one is to use a pre-trained CNN to recognize every whole target, but it doesn't work well, because numbers of figures on targets can change(like 3, 23, 422, etc.), without a fixed length, CNN can't recognize targets easily; another difficulty is that there can be thousand of results possible, because there are figures with decimal point. If we want to recognize all possible figures, we have to train CNN to learn features of numbers from 0.1 to at least 200.0, which places high demands on the accuracy and complexity of the neural network. With all these difficulties, we can only reach a final accuracy of 5% - 20%.

The second idea is that we separate the whole target into single characters(single figure), then use CNN to recognize these single characters. By doing so, we can just train CNN to learn features of figures from 0 to 10. So we firstly use Fast Line Detector to find out the direction of characters and rotate them to horizontal position, then split each character by using projection layout segmentation algorithm. In the meanwhile we detect also if there is a decimal point by using FloodFill function who can count the area of interconnected domains. In practice we can nearly 100% split characters correctly. The accuracy of the detection of decimal points is also high, around 80% to 95%. Once we get all single characters, we use a pre-trained CNN to recognize them, and CNN works great for single character recognition, it can reach an accuracy of 90% with a very simple training.

4. Final performance of the project

Eventually we chose the second idea, the total accuracy of whole processes is 24%-85%, calculated by positioning accuracy (60%-90%) * detection of decimal point accuracy (80%-95%) * segmentation-recognition-combination accuracy (51%-100%). The reason that the total accuracy varies from 24% to 85% is that for certains maps, there are lots of noises connected with figures, which makes positioning and segmentation steps very difficult and the accuracy of the recognition depends on the preparation of the training set of the neural network.

5. Short conclusion

The technical coverage of this project is relatively large, including image processing, datasets building and neural networks. The progress and results of the project are in line with expectations, but there are still some aspects which can be improved.