Identification of geographical names in scanned maps using OCR and ML

Background

In 2023 the Norwegian Mapping Authority assigned a group of bachelor students to a proof-of-concept project aiming to use machine learning (ML) to identify geographical names in scanned maps. After a promising pilot trial the project was continued by interns.

Source data

Initially, the input data was a selection of maps covering the town of Voss, mainly editions from the M711 1:50 000 map series produced between 1953-2006. The second group expanded the scope to include various map series covering the whole of Norway, 9425 maps in total.

Choice of ML-model

After assessing several models, the group chose Google Cloud Vision Text Detection.

Work flow

The workflow can be broken down to three stages:

Preprocessing: Images of scanned maps were preprocessed using several techniques such as rescaling, thresholding and denoising.

Inference: The model's predictions were annotated with a bounding box and a character string representing the interpretated text.

Postprocessing: Results (predictions) were processed and written to a PostGIS database. Each entry in the database contains the interpreted text (geographical name), source map metadata, bounding boxes in image and real-world coordinates.

User interface

The user interface (UI) providing a text search and basic web map capabilities, viewable in any web browser. Search results are presented with a position marker in the map window, text, source information and an image of the original map.

Assessment

Although it is difficult to provide a specific percentage, the ML model was able to identify many geographical names from the maps. The database has extracted over 4 440 000 geographical name instances from 9425 maps, although this number does not represent unique names since the same geographical name will often be present on various maps covering the same area. For

comparison, the Norwegian Place Name Registry (SSR) contains >1 150 000 geographical names.

The UI has proved an efficient way to search for geographical names in case handling, where documentation is a time-consuming task. It could also be of interest to academia and the public.

Potential

Although the project is concluded, the concept has further potential. The percentage of names identified by the ML model could be higher with a specialized model. Other potential data sources are text from books, gazetteers or academic archives which often have a relative position reference such as map grid position which could be converted to real world coordinates.