

Introduction

This master’s thesis focuses on automating data extraction from the historical Stable Cadastre using neural networks. The aim of the thesis was the digitization of the Stable Cadastre, specifically the detection of parcel numbers from cadastre maps and text recognition from written records. The primary challenge lies in the materials’ age and degraded condition. To address this, two object detection and four text recognition methods were utilized.

Methodology & Analysis

As these are historical documents, the primary challenge is their age: the papers are yellowed, stained, and hand-written. Furthermore, the written records suffer from various geometric deformations. Based on an analysis of object detection methods, the EAST and YOLO architectures were selected. For text recognition, both classical architectures based on Convolutional Neural Networks (CNN) and modern Large Vision-Language Models (VLM) were considered. Four architectures were subsequently selected for implementation: Kraken, GOT-OCR 2.0, HunyuanOCR, and QWEN3.

Key Aspects of Implementation

A crucial part of the implementation was the development of two annotation applications, one for parcel numbers on maps and another for table cells in written records. Of note is a custom grid-detection algorithm that first uses a small YOLO model to detect line intersections, then filters and clusters the results into rows and columns. This algorithm effectively processes tables with horizontal misalignment and curved rows. A detailed schema of the algorithm is shown in Figure 1, and the final detection is shown in Figure 2. Another phase involved hyperparameter optimization and subsequent training for both detection and text recognition models.

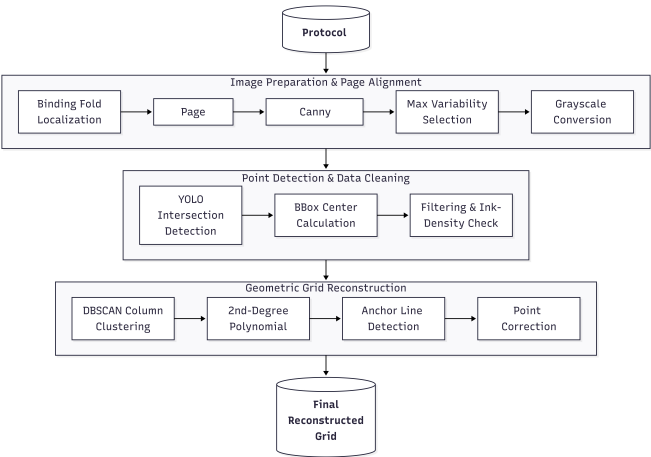


Figure 1: Grid detection algorithm schema.

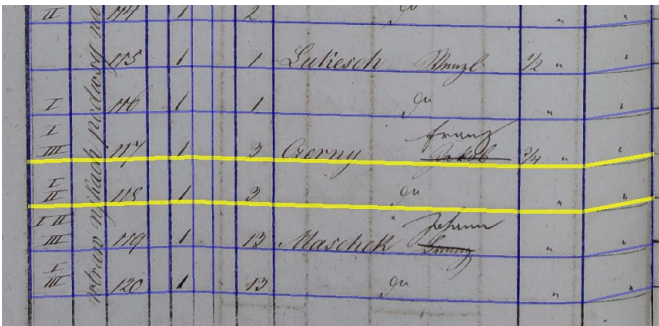


Figure 2: Example of grid detection (selected yellow row).

Results

Experiments demonstrated that high model reliability is achievable despite the age of the documents. In detection, the EAST model performed best, achieving an impressive Recall rate of 94.0%, meaning it can detect up to 94.0% of parcel numbers on the map. The main limitation is densely built-up areas, where numbers are tightly packed and difficult to detect. Detection results are shown in Table 1.

Model	Precision	Recall	F1 score	mAP50	IoU
EAST	91.0	94.0	92.5	85.9	75.0
YOLO	86.4	90.0	88.2	87.3	79.5

Table 1: Detection results [in %].

In both parcel-number and written-record recognition, the Hunyuan model achieved excellent results. In parcel number recognition, the GOT model is also noteworthy; it is up to three times faster than Hunyuan while achieving comparable performance. Surprisingly, the Kraken model achieved results comparable to those of the large VLM model, QWEN. Otherwise, in written-record recognition, the GOT and Kraken models couldn’t match the performance of large VLM models. The results are shown in Table 2.

Metric	Kraken	GOT	Hunyuan	QWEN
Parcel number				
CER	20.7	17.0	16.0	20.4
Accuracy	61.4	66.7	67.4	64.2
Time	0.1s	0.7s	2.2s	4.3s
Written record				
CER	31.8	21.2	8.6	11.8
Accuracy	0.3	0.6	27.0	11.4
Correct Column Count	48.6	81.5	92.6	95.2

Table 2: Results of OCR models [in %].

Conclusion

Despite the document’s age, the models achieved high accuracy and low error rates. The trained models were successfully integrated into the annotation applications, resulting in a functional system capable of automatically processing historical Stable Cadastre maps and written records. This work confirms that modern neural networks can effectively replace manual data extraction from historical documents, facilitating further work with historical data in the Czech Republic.