Introduction

Toponym Detection

Toponym Classification

Experiments

Conclusions

Historical Map Toponym Extraction for Efficient Information Retrieval

Ladislav Lenc, Jiří Martínek, Josef Baloun, Martin Prantl, Pavel Král

Dept. of Computer Science & Engineering
University of West Bohemia
Plzeň, Czech Republic

NTIS - New Technologies for the Information Society
University of West Bohemia
Plzeň, Czech Republic

llenc,jimar,balounj,perry,pkral@kiv.zcu.cz

May 23, 2022

Introduction I

Task

- Information Retrieval (IR) in historical hand-drawn maps
- Two types of map toponyms:
  1. **Municipal toponyms (printed)**
     (names of towns, municipalities, villages, ...)
  2. **General toponyms (handwritten)**
     (road names, forrest, hills, ...)
- Automatic processing of map toponyms (place names):
  - Toponym detection;
  - Toponym classification;
  - Toponym text recognition (OCR);
- Toponyms used as keywords in users queries in IR system

Ladislav Lenc, Jiří Martínek, Josef Baloun, Martin Prantl, Pavel Král

Introduction
Toponym Detection
Toponym Classification
Experiments
Conclusions

Historical Map Sheets

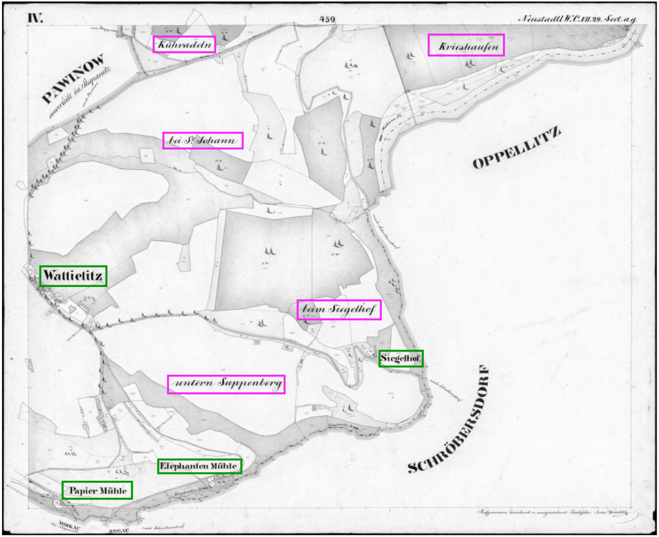

Figure : Map sheet with highlighted toponyms.

Historical Maps Sheets

- Scanned map sheets from the 19th century
- Austro-Hungarian Empire teritory
- Maps covers the area of the current Czech Republic but toponyms in German language
- 800 map sheets and 2900 annotated toponyms;
- dataset available for research purposes[1]

Table : Numbers of handwritten and printed toponyms within our dataset.

Dataset	Map Sheets	Handwritten Toponyms	Printed Toponyms
Train	650	2050	335
Test	100	305	41
Dev	50	141	28

[1] `https://corpora.kiv.zcu.cz/nomenclature/`

Overall System

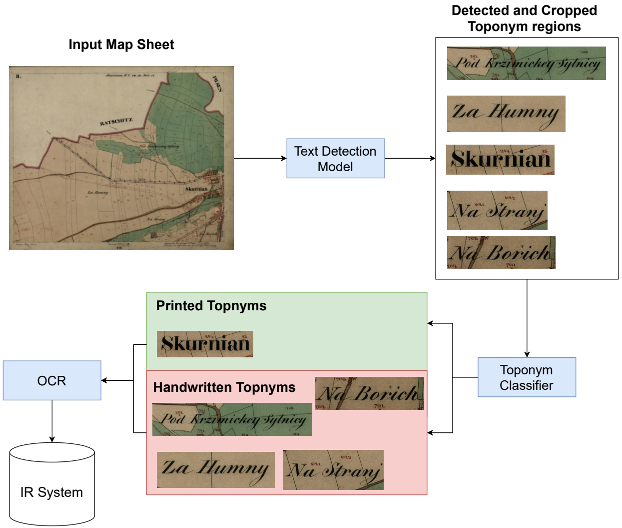

Figure : Overall Processing Pipeline

Toponym Detection

- Baseline approach based on thresholding, morph. operations and connected component analysis (CCA)
- We compared and evaluated several text detection models:
  - **HP-FCN**: High Performance Fully Convolutional Network
  - **EAST**: an efficient and accurate scene text detector
  - **Faster R-CNN**
  - **YOLOv5**
- YOLOv5 and Faster R-CNN capable of classification

Ladislav Lenc, Jiří Martínek, Josef Baloun, Martin Prantl, Pavel Král

Introduction
Toponym Detection
Toponym Classification
Experiments
Conclusions

Experiments – Toponym Detection

Table : Results on 0.5 IoU level (Avg AP: interval 0.5 – 0.95 with 0.05 step)

Model	IoU@50 Prec.	Rec.	F1	AP	Avg AP
CCA (baseline)	19.5	60.4	29.5	11.3	2.78
EAST Detector	84.5	**89.9**	**87.1**	**77.8**	**46.7**
HP-FCN	65.4	75.4	70.1	44.4	20.6
YOLOv5	84.6	79.2	81.8	67.1	37.1
Faster R-CNN	**87.2**	80.9	83.9	71.2	41.8

Table : Results on 0.75 IoU level

Model	IoU@75 Prec.	Rec.	F1	AP
CCA (baseline)	10.7	33.1	16.2	0.27
EAST Detector	77.5	**82.4**	**79.9**	**51.3**
HP-FCN	53.9	62.2	57.8	17.1
YOLOv5	76.4	71.7	73.9	39.7
Faster R-CNN	**80.6**	75.0	77.7	45.4

Toponym Classification I

- Input = cropped images from the previous step
- Pre-processing – noise reduction, binarization, CCA
- Toponym classification algorithm based on KAZE image descriptors (inspired by writer identification [1])

[1] Xiong, Y.J., Wen, Y., Wang, P.S.P., Lu, Y.: *Text-independent writer identification using sift descriptor and contour-directional feature*. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015

Toponym Classification II

Codebook generation

- Based on training set
- KAZE is applied $\rightarrow$ set of descriptors;
- Descriptors are clustered with K-means $\rightarrow$ 100 centroids;

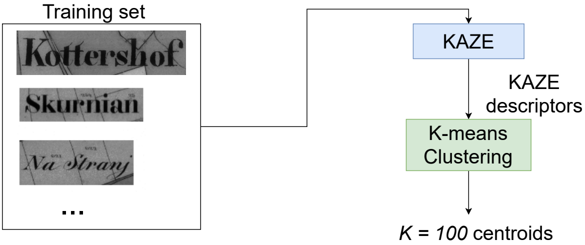

Figure : Codebook generation

Toponym Classification III

Image Representation

- Various number of desciptors are produced for input image
- Histogram of the closest centroid is associated with a label

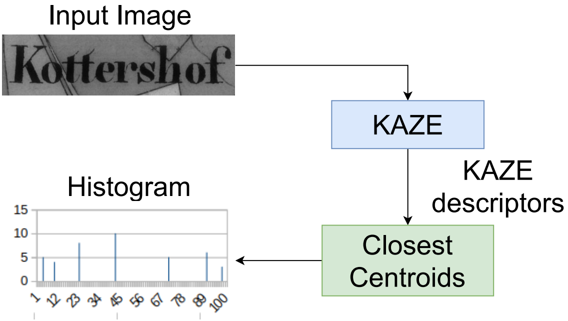

Figure : Image Representation

Toponym Classification IV

Toponym Prediction

- Image features = Histogram of closest centroid
- Find *N* nearest histograms
- We predict the majority class occurring in the *N* most similar histograms

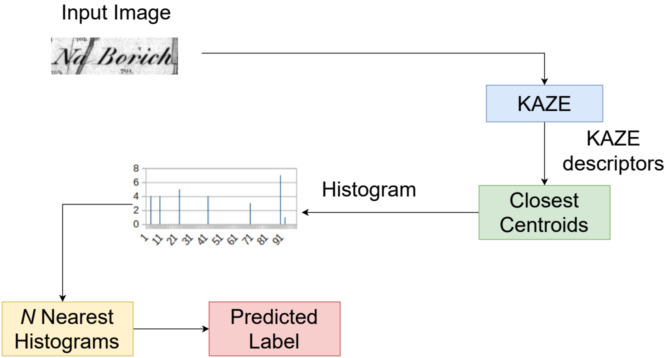

Figure : Prediction phase

Ladislav Lenc, Jiří Martínek, Josef Baloun, Martin Prantl, Pavel Král

Introduction
Toponym Detection
Toponym Classification
Experiments
Conclusions

Experiments – Toponym Classification

- YOLOv5 and Faster R-CNN classification results slightly worse results
- Our approach has comparable results for all detection methods
- Robust method applicable for different sizes of the detected region

Table : Toponym classification results; accuracy (ACC) in %

Detection Approach	**Classification Approach**	**ACC**
CCA (baseline)	Proposed	98.7%
EAST	Proposed	99.1%
HP-FCN	Proposed	**99.2**%
YOLOv5	Proposed	98.8%
Faster R-CNN	Proposed	98.8%
YOLOv5	YOLOv5	97.6 %
Faster R-CNN	Faster R-CNN	98.2 %

Experiments – OCR

- Baseline OCR: Tesseract ENG
- Trained Tesseract model for both Printed and Handwritten toponyms
- Character Error Rate (CER) on 346 bounding boxes from Test toponyms
- Combined Tesseract = pick $Tess_P$ or $Tess_H$ based on toponym classification predictions

Table : OCR Results with Tesseract

	Printed	**Handwritten**	**All**
Number of Toponyms	41	305	346
Tesseract ENG (baseline)	0.153	0.477	0.437
$Tess_P$ (trained)	**0.061**	0.512	0.459
$Tess_H$ (trained)	0.076	**0.185**	0.185
Combined Tesseract	–	–	**0.171**

Conclusions I

- **Toponym Detection**
  - EAST model has the best average precision values
  - HP-FCN worse results than other models
- **Toponym Classification**
  - Our Toponym classification algorithm better performance (99%)
  - Small amount of training examples is sufficient for reasonable results
- **OCR**
  - Trained Tesseract $\rightarrow$ significant improvement (17% CER)
  - Information about toponym class valuable $\rightarrow$ pick the specialized *tessdata*

Conclusions II

- Faster R-CNN and YOLOv5 obtained very good detection and classification results
- The best strategy: **separated training**
- Map sheets are currently processed and our toponym extraction approach is deployed

Future Work

- Error Correction method
- More types of toponyms $\rightarrow$ distinguish between cadastres, rivers, hills, etc.
- Deployment of our approach on map sheets from different era

Acknowledgements

This work has been partly supported from ERDF "Research and Development of Intelligent Components of Advanced Technologies for the Pilsen Metropolitan Area (InteCom)" (no.: CZ.02.1.01/0.0/0.0/17_048/0007267) and by Grant No. SGS-2022-016 "Advanced methods of data processing and analysis".