The Pinkas Dataset

Berat Kurar, Jihad El-Sana

Department of Computer Science,
Ben-Gurion University of the Negev,
Israel

Irina Rabaev

Software Engineering Department,
Shamoon College of Engineering,
Israel

Introduction

- Benchmark datasets are important for evaluation and comparison of different methods
- We introduce the Pinkas dataset – the first dataset in medieval handwritten Hebrew:
    - 30 pages manuscript with its ground truth at page, line and word level
    - baseline experiments with three methods for word spotting

The Dataset Description

- Records of Frankfurt community, dated 1500 -1800
- Mixture of medieval Hebrew
- Different handwritings, writers were not professional scribes
- Numerous degradation types
- Complex layout
- Challenging both for computer and human analysis

Annotation

- The dataset is annotated at page, line and word level using Aletheia system [1]
- The ground truth is in PAGE format [2]
- The initial annotation was corrected by a paleographer expert

Main Text	Side text	Signature marks	Dates
108	7	13	11

Number of regions per category

Lines	Words	Word classes
1013	13744	3387

Total amount of lines, words and word classes

Word Spotting Experiments

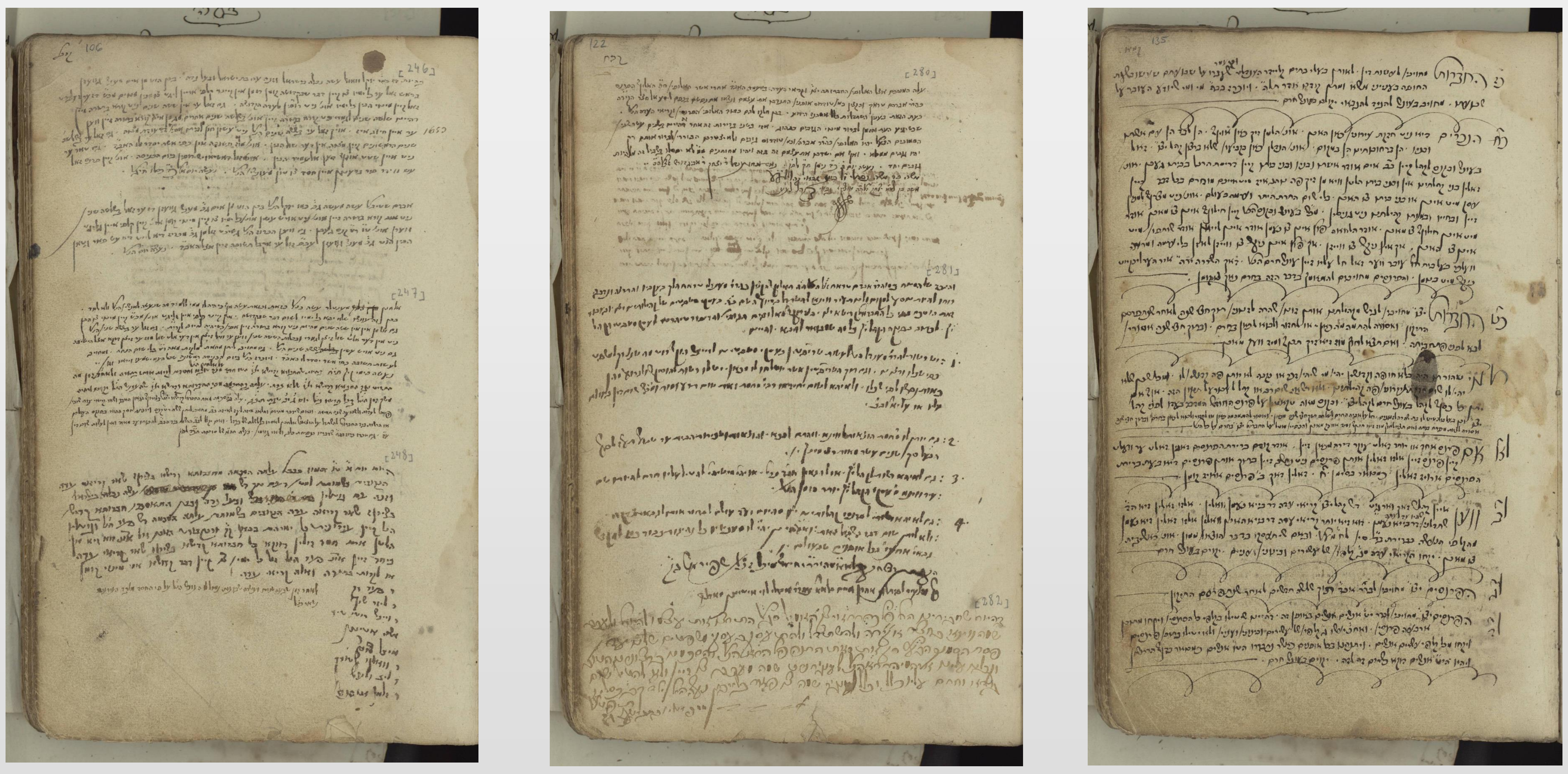
- Official train and test sets
    - Train set - first 24 pages (80%)
    - Test set - last 6 pages (20%)
- Three different methods
    - Siamese CNN [3]
        - supervised segmentation-based
    - PHOCNet CNN [4]
        - supervised segmentation-based
    - Exemplar SVM [5]
        - unsupervised segmentation-free

Train		Test		
Classes	Samples	Classes	Samples	OOV
3117	10397	1251	3278	603

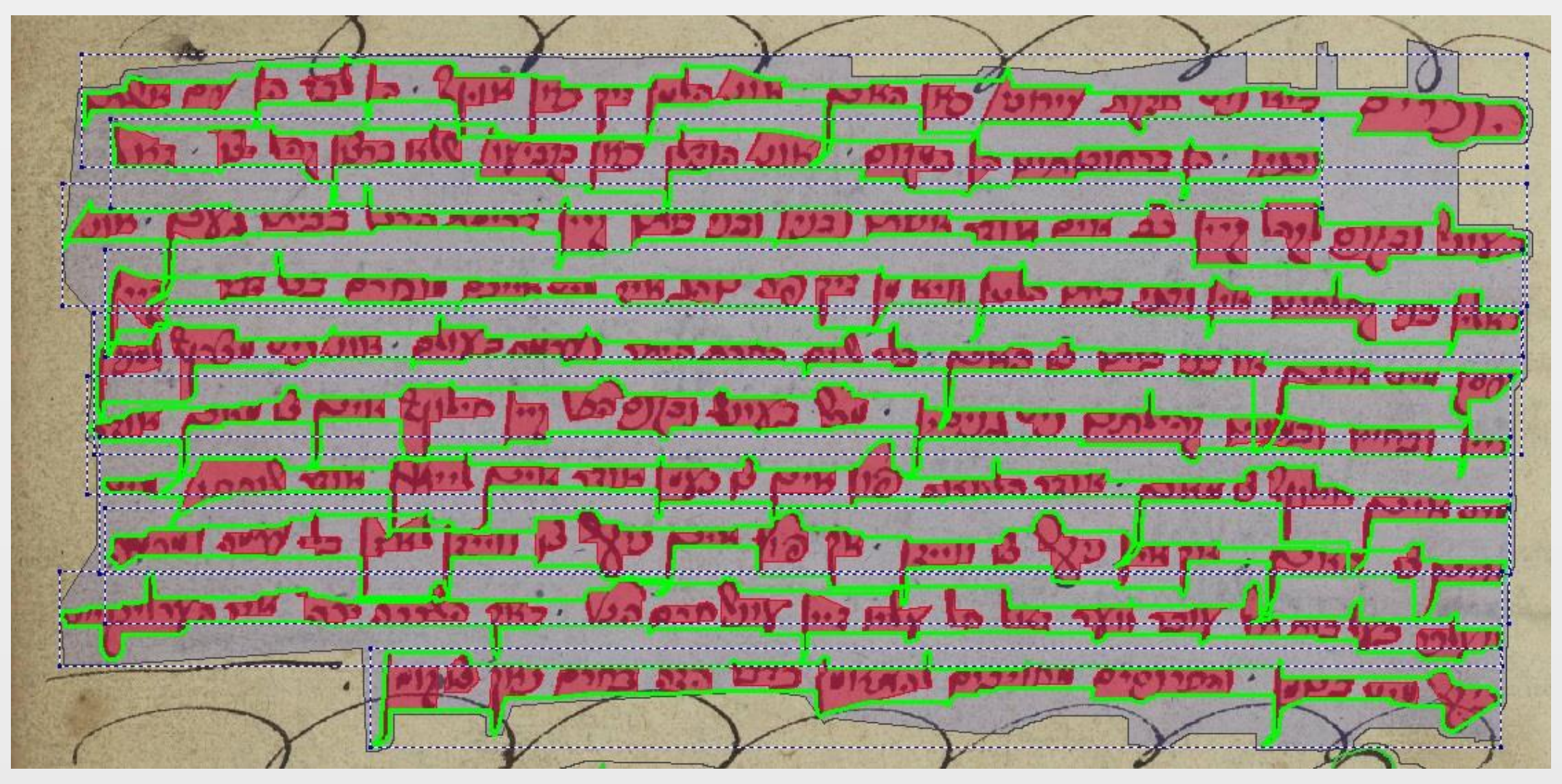
Statistics on train and set partition

Siamese CNN	PHOCNet	PHOCNet One hot	Exemplar SVM
61.5	56.6	53.3	1.5

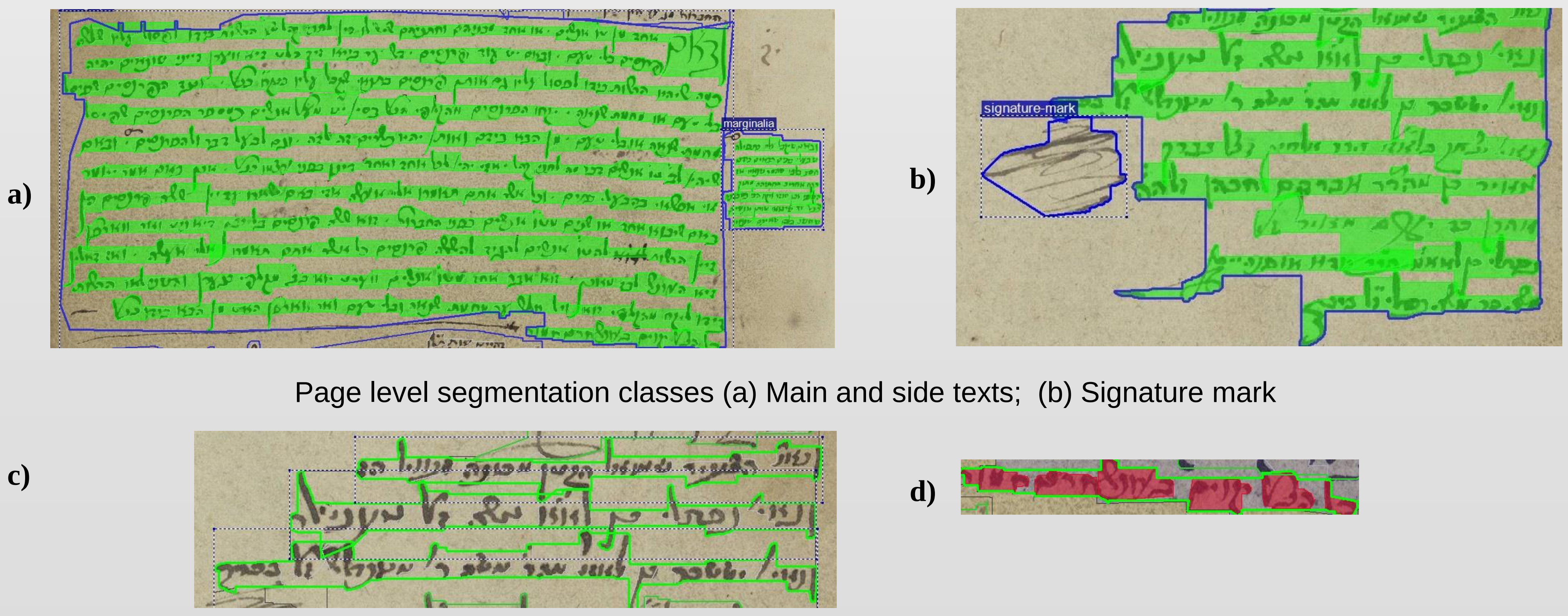
MAP results of word spotting methods



Sample document images from the Pinkas dataset. Paragraphs are separated by drawings or by space. Some paragraphs are assigned by a number which is written in a spatial proximity to them.



Main text (purple), line (green) and word (red) segmentation levels



Page level segmentation classes (a) Main and side texts; (b) Signature mark

Text line and word segmentation (c) Segmentation of overlapping lines; (d) Word segmentation. Note there is no space separation between the fifth and the sixth words

Conclusions

- The Pinkas dataset is a challenging dataset which contributes to the diversity of benchmarking standards
- Ground truth at page, line and word level
- An official train and test set partition is defined
- Baselines are set by three word spotting methods
- The results show that there is a big room for improvement
- Currently, the dataset is available for download at: https://www.cs.bgu.ac.il/~berat

Future directions

- In future research we plan to run baseline experiments for page and text line segmentation
- We are currently extending the dataset and are going to make it available

Primary references

[1] C. Clausner, S. Pletschacher, and A. Antonacopoulos, “Aletheia - an advanced document layout and text ground-truthing system for production environments,” in 2011 International Conference on Document Analysis and Recognition. IEEE, 2011, pp. 48–52.

[2] S. Pletschacher and A. Antonacopoulos, “The page (page analysis and ground-truth elements) format framework,” in 2010 20th International Conference on Pattern Recognition. IEEE, 2010, pp. 257–260.

[3] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, “Signature verification using a “siamese” time delay neural network,” in Advances in neural information processing systems, 1994, pp. 737–744.

[4] S. Sudholt and G. A. Fink, “Phocnet: A deep convolutional neural network for word spotting in handwritten documents,” in 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE, 2016, pp. 277–282.

[5] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, “Efficient exemplar word spotting.” in Bmvc, vol. 1, no. 2, 2012, p. 3.