The HHD Dataset

Irina Rabaev, Alexander Churkin

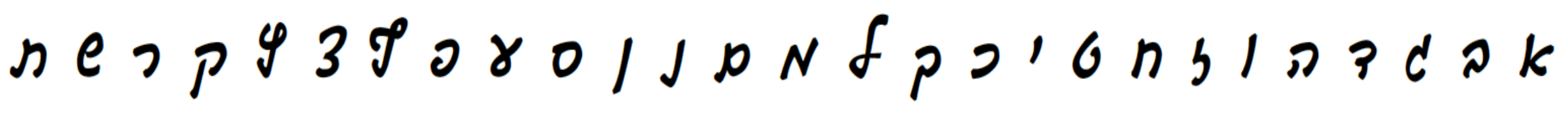
Software Engineering Department
Shamoon College of Engineering
Israel

Berat Kurar Barakat, Jihad El-Sana

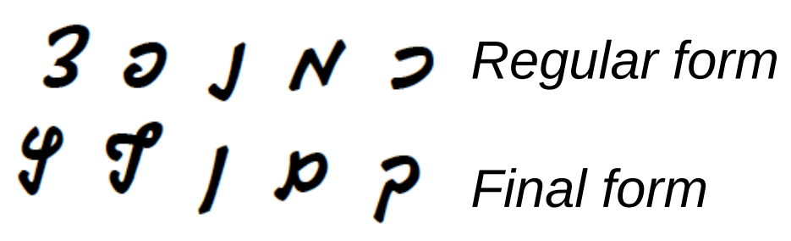
Department of Computer Science
Ben-Gurion University of the Negev
Israel

Introduction

- Benchmark datasets are important for evaluation and comparison of different methods
- We introduce the HHD dataset – a handwritten Hebrew dataset
  - 1000 scanned handwritten forms
  - ground truth at text line, word and character levels
  - baseline experiments for initial small subset, HHD_v0
- Hebrew alphabet consists of 22 consonant only letters



  - five letters have additional final form
  - high visual similarities among letters



Groups of very similar letters

The Dataset Description

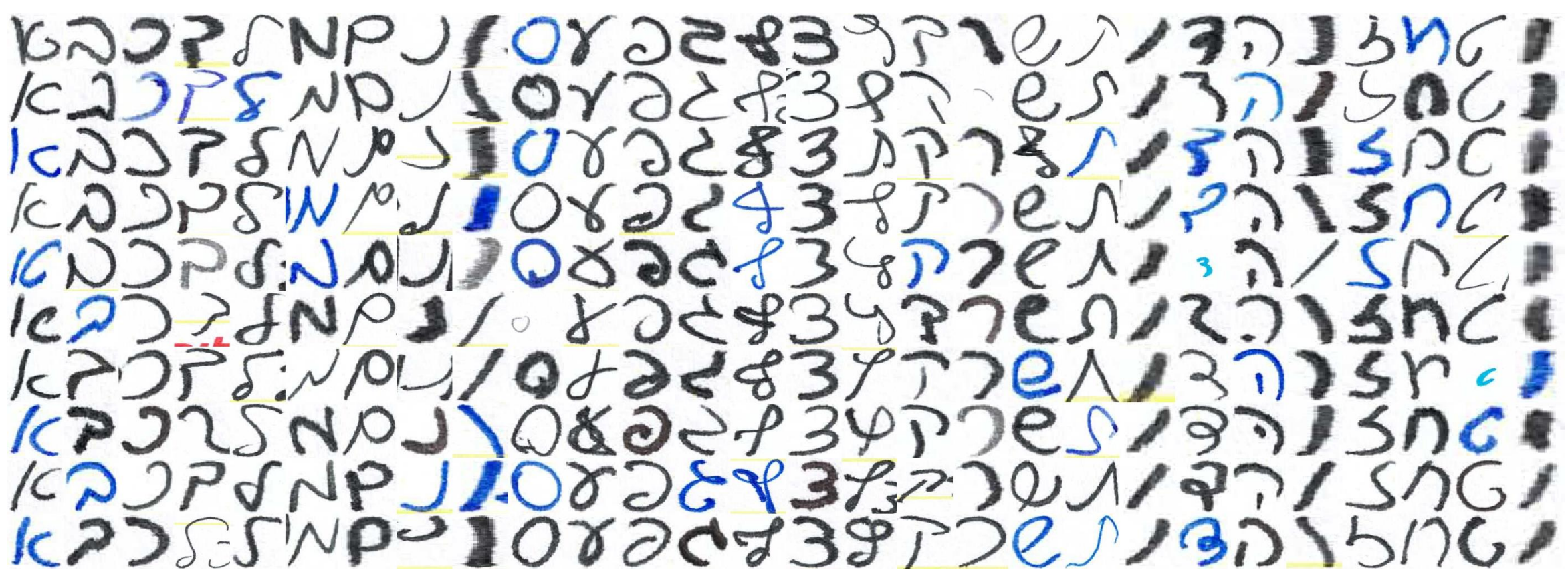
- Scanned handwritten forms (600 dpi, TIFF format)
  - filled by individuals from different age groups and educational backgrounds (from as young as 11 years old and as old as late 40s)
  - no restriction on pen/pencil type or color
- 63 variations of the forms
  - forms A - M contain isolated sentences and words
    - forms A - E are based on pangrams
  - forms 1 - 50 contain text paragraphs from four categories
    - general news
    - scientific articles
    - children's books
    - economy news

Annotation

- The structure of the forms facilitates automatic ground truth generation
- The ground truth is in PAGE format [1]
- The initial annotation is verified and corrected by a human

Initial Experiments

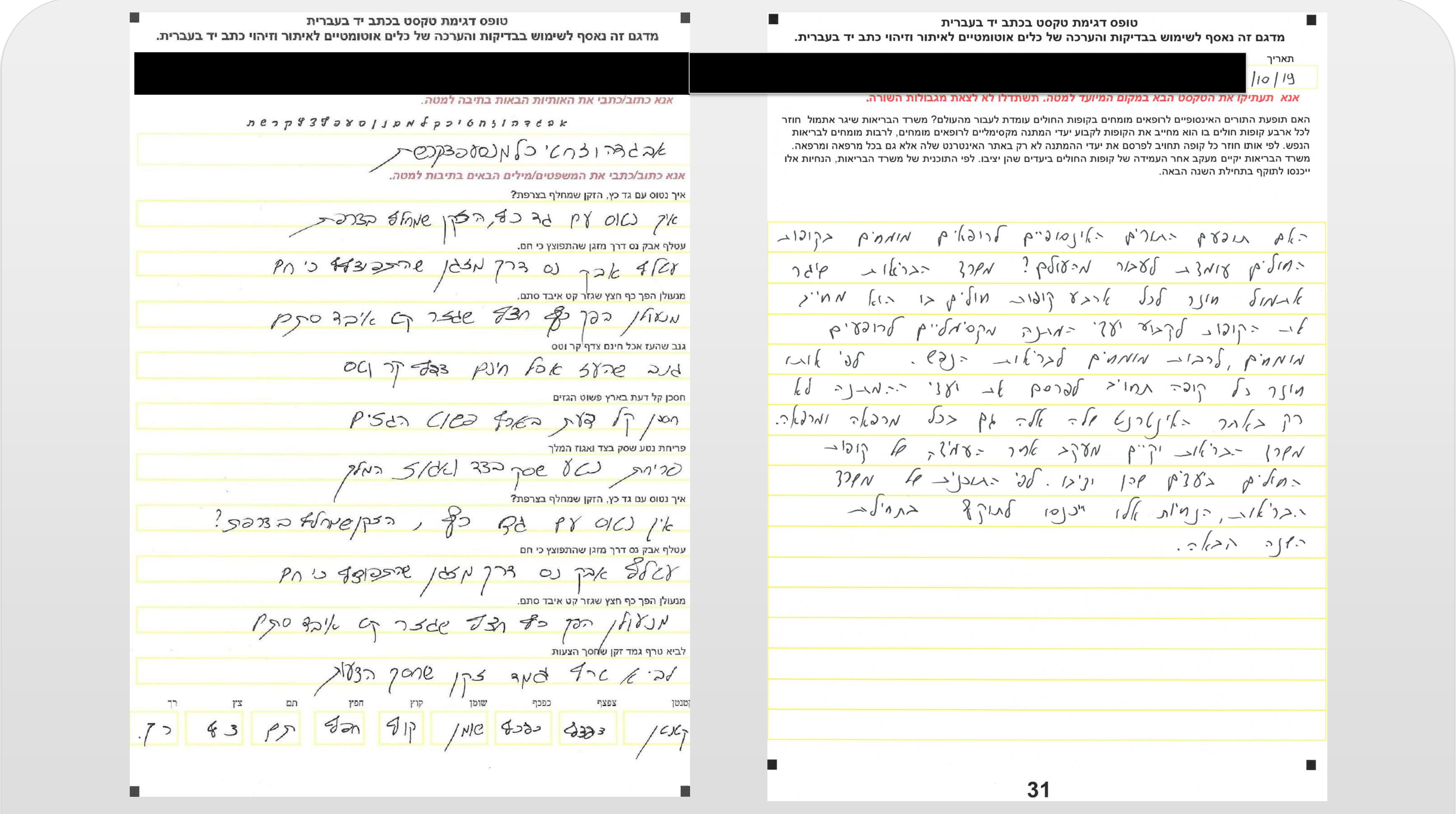
- Character classification on initial subset, HHD_v0
  - 5099 isolated character images divided into train (3965 images) and test (1134 images) sets



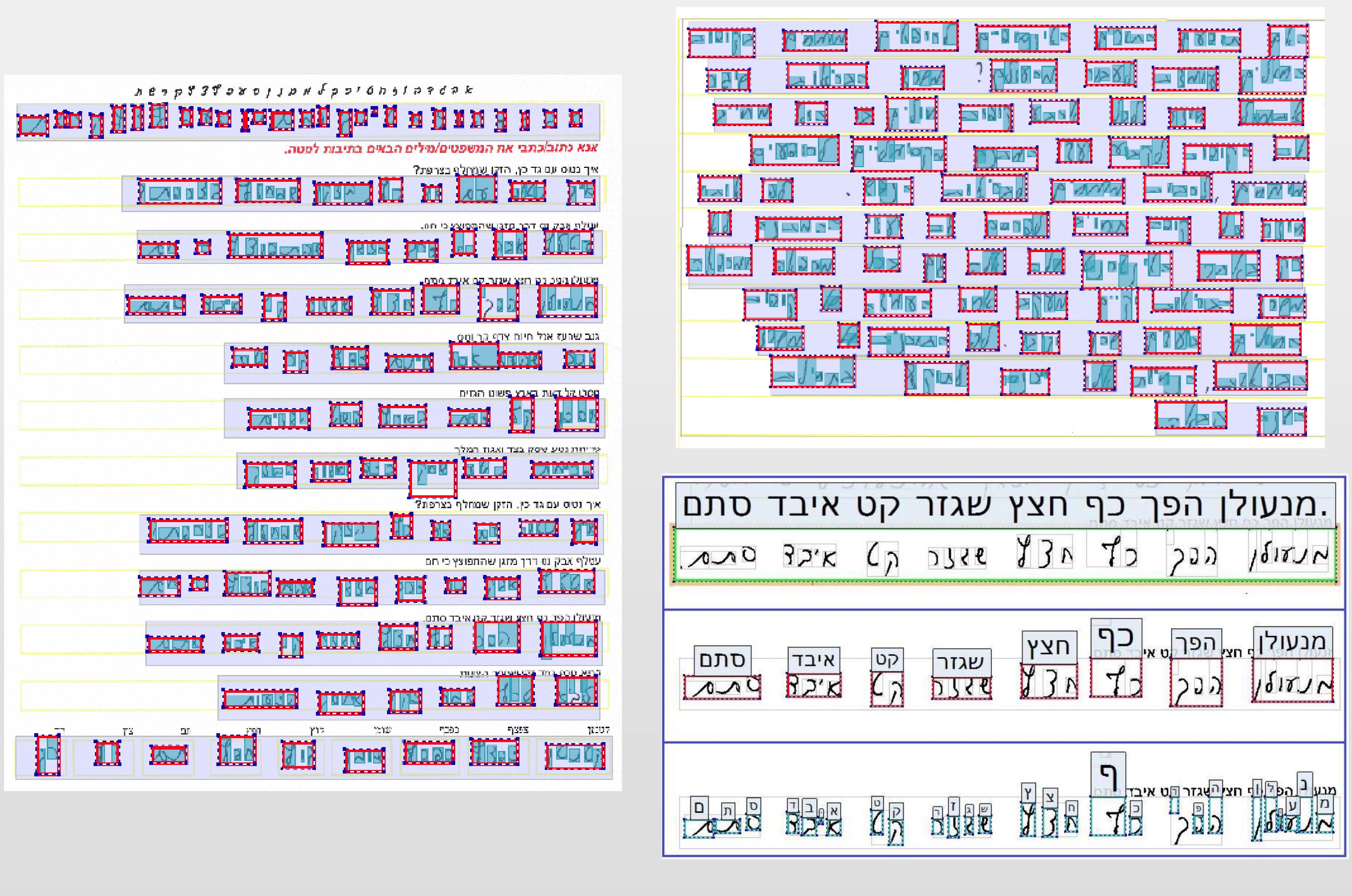
- Three different CNNs
  - Simple CNN with three hidden layers
  - AlexNet [2]
  - ResNet [3]

	Train accuracy	Test accuracy
Simple CNN	96.62	72.57
AlexNet	99.55	78.21
ResNet	**100**	**84.9**

Character classification results on HHD_v0


[redacted]
[redacted]

Examples of different forms: sentences based on pangrams (on the left), and a text paragraph (on the right). The text boxes in the first row of each form are provided to fill the personal information and are blacked out. The black rectangles at the corners of the page are used for aligning.



Annotated handwritten forms with superimposed ground-truth.

Conclusions

- The HHD is a dataset of modern handwritten Hebrew document images
- Ground truth is generated at text line, word and character levels
- An initial subset HHD_v0 of the dataset is available for download at: https://www.cs.bgu.ac.il/~berat/data/hhd_dataset.zip
- Baselines for character classification on HHD_v0 are set by three different CNNs

Future directions

- We are currently extending the dataset and going to make it publicly available
- Future plans include
  - Initial experiments for word spotting, text line alignment and writer verification
  - Applying cross-domain transfer learning: use of networks that have been pre-trained on HHD for historical document images

Primary references

[1] 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.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778