WORDSPOTTING FOR HISTORICAL DOCUMENTS

Leonard Rothacker, Sebastian Sudholt, Eugen Rusakov, Matthias Kasperidus and Gernot A. Fink

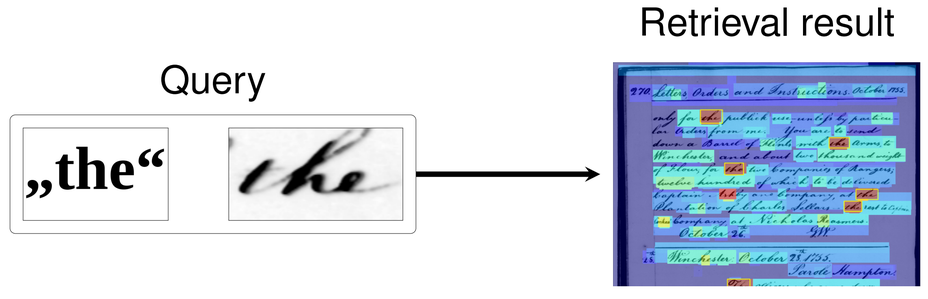
Department of Computer Science, TU Dortmund University, Germany

technische universität dortmund

Abstract

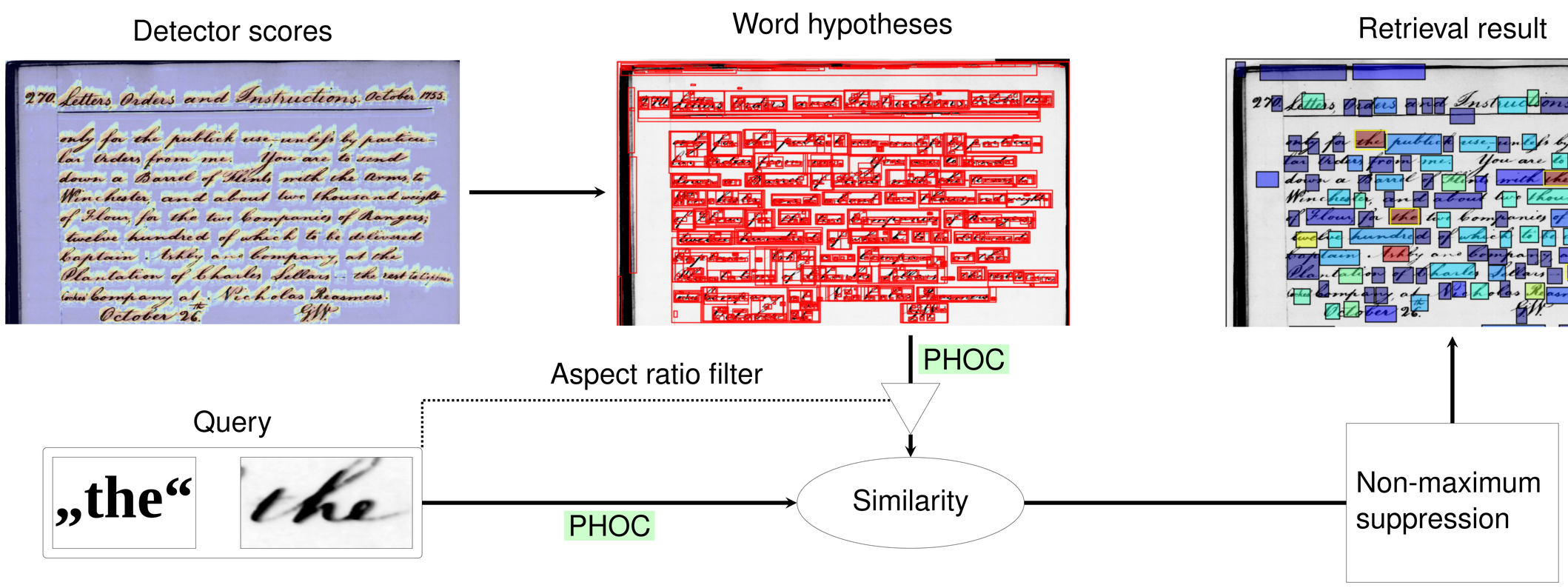
The generation of word hypotheses for segmentation-free word spotting on document level is usually subject to heuristic expert design. This involves strong assumptions about the visual appearance of text in the document images. In this paper we propose to generate hypotheses with text detectors. In order to do so, we present three detectors that are based on SIFT contrast scores, CNN region classification scores and attribute activation maps. The uncertainty in the detector scores is modeled with the extremal regions method. Retrieving word hypotheses is based on PHOC representations which we compute with the PHOCNet. We evaluate our method on the George Washington dataset and the ICFHR 2016 KWS competition benchmarks.

Wordspotting



The generation of word hypotheses for segmentation-free word spotting on document level is usually subject to heuristic expert design. This involves strong assumptions about the visual appearance of text in the document images. In this paper we propose to generate hypotheses with text detectors. In order to do so, we present three detectors that are based on SIFT contrast scores, CNN region classification scores and attribute activation maps.

Method



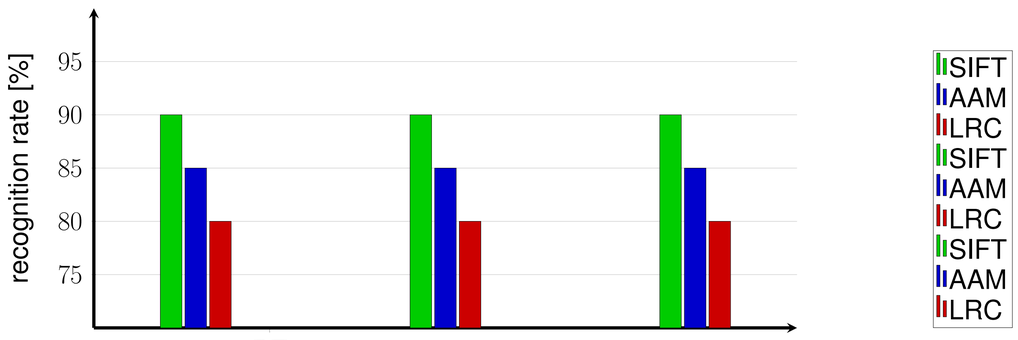
Results

Estimating orientation

Method	Georg Washington DR	mR	mAP
2-PHOCNet System	96.91	83.04	55.57
1-PHOCNet System	**97.35**	**90.96**	**77.73**
AttributeSVM [?]	86.49	21.71	–

Number of correctly oriented postcards after the pre-processing on a validation set of 256 postcards. 65.8% of the images in the validation set are landscapes and thus already correctly oriented without any processing.

Recognition of thematic categories



Classification results for different features (cf. [3]) and feature ensembles on the 256 postcard dataset and a large unexplored set with more than a thousand postcards (recognition rates could only be estimated as no ground truth is available).

References

[1] Bley, B.: Feldpostkarten im 1. Weltkrieg (Feldpost Postcards of WWI). *Private Collection*

[2] Brocks, C.: Die bunte Welt des Krieges: Bildpostkarten aus dem Ersten Weltkrieg 1914-1918 (The Colorful World of the War: Picture Postcards from the First World War 1914-1918). Klartext-Verlag, Essen (2008), (in German)

[3] Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: *Computer vision and pattern recognition (CVPR)*, 2010