Word Hypotheses for Segmentation-free Word Spotting in Historic Document Images

Leonard Rothacker, Sebastian Sudholt, Eugen Rusakov, Matthias Kasperidus and Gernot A. Fink
Department of Computer Science
TU Dortmund University
44227 Dortmund, Germany
Email: {firstname.lastname}@tu-dortmund.de

***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 TPP-PHOCNet. We evaluate our method on the George Washington dataset and the ICFHR 2016 KWS competition benchmarks. In the evaluation we show that high word detection rates can be achieved. This is a prerequisite for high retrieval performance that is competitive with the state-of-the-art.**

I. Introduction

Word spotting is an efficient method for making document images searchable. Therefore, it provides an essential functionality for working with large document image collections. The approach is efficient since the search functionality is directly implemented and not a by-product of a more complex task, typically transcription. Most commonly, the search query is either given as a word image in query-by-example scenarios or as text in query-by-string scenarios. All word spotting methods need to either explicitly (segmentation-based) or implicitly (segmentation-free) segment the document collections into word image hypotheses. State-of-the-art methods project the word images into an embedded attribute space [1] using *Convolutional Neural Networks (CNN)* [2], [3]. In this space, word spotting can then be accomplished through a simple nearest neighbor search. For historic documents, automatic segmentation is especially challenging due to high variability in writing style, document layout, visual appearance of ink and paper, as well as aging artifacts.

Segmentation methods that have been successful in modern document images, such as projection profiles or connected components, are likely to fail for historic documents. Instead, these methods have to be manually tuned to the document collection's specificities. Interesting segmentation methods have been presented in [4] and [5]. Within the scale space approach in [4], some parameters can automatically be derived from data. The approach in [5] uses a CNN for classifying segmentation hypotheses. The visual word appearance is, therefore, learned from annotated sample data. However, methods addressing solely segmentation need to detect words without recognizing them or, in case of word spotting, without taking relevance to the query into account. Therefore, these methods have to rely on discriminative characteristics of the document collections considered. In the challenging scenario of historic document images, it remains questionable, if suitable characteristics can automatically be extracted. This aspect can potentially limit the generalization capability.

In order to be more robust with respect to word size variability, our segmentation-free word spotting method is inspired by approaches using local text detectors. In many cases text detectors are solely built on connected components, e.g. [6]–[8]. This has two important drawbacks. First, the detectors are dependent on document image binarization. In historic document images binarization is difficult due to fading ink, low contrast and inhomogeneous backgrounds. This makes detections imprecise. Second, it can be difficult to derive word hypotheses from connected components. Since connected components can represent parts of words, single words or multiple words, heuristic strategies for combining connected components are required.

For these reasons, we propose to generate word hypotheses based on higher-level feature representations that indicate word occurrences. First, we predict scores for certain document image regions. These scores reflect whether the respective region contains text or not. The uncertainty of these scores is then explicitely modeled with extremal regions (ERs) [9] that have been very successful for text detection in natural scene images, cf. [10]. The ER approach generates hypotheses of word bounding boxes. For these, PHOCs are predicted using a TPP-PHOCNet [2]. This is essentially a *Region-based CNN (R-CNN)* [11] framework. After predicting the PHOCs, word spotting can be performed through a nearest neighbor search.

Generating the local text scores is a critical part of our method. Here, we consider three different approaches: SIFT contrast scores, local region classification scores generated with a CNN and local word region scores obtained with an extension of CNN class activation maps [12].

Sec. II presents segmentation-free word spotting methods and briefly reviews extremal regions. Our segmentation-free word spotting approach and its evaluation are presented in Sec. III and Sec. IV. Finally, conclusions are drawn in Sec. V.

II. Related Work

Word spotting methods that are addressing segmentation and retrieval jointly are referred to as segmentation-free. In order to address the segmentation problem at document level, mainly two different approaches can be identified. Based on local text detectors, different competing word hypotheses are obtained, cf. [3], [6]–[8], [13]. In contrast, patch-based approaches densely sample word hypotheses from the document images, cf. e.g., [14]–[18]. By searching the full document, patch-based approaches do not rely on heuristic detectors. However, they limit the search to a single patch size per query, thus assuming that the size variability is relatively low. Finally, in both approaches word hypotheses are ranked according to similarity with the query and overlapping hypotheses are suppressed if they obtained a non-optimal score. Segmentation-free methods, therefore derive the segmentation during the retrieval process and do not rely on a given segmentation that is assumed to be correct.

Segmentation-free word spotting based on PHOC representations, cf. [1], has been presented in [18] and [8] for the first time. Here, the document is divided into a number of blocks and a PHOC is predicted for each block. For efficient patch-based retrieval, an integral image over the block-wise PHOC vectors is computed. In order to improve the results, a regression is learned which projects PHOCs and predictions into a common subspace. At query time, the query PHOC is projected into this subspace. The similarity between the query and the patches is then determined through a dot product. While all patches are considered in [18], the approach presented in [8] adds an indexing stage in order to efficiently detect regions of interest. In this stage, connected components in close proximity to each other are combined in order to obtain word hypotheses. For retrieval, candidate word regions, obtained from the index, define the document image search area for the patch-based framework presented in [18]. For query-by-example [18] the patch size equals to the size of the query word image and for query-by-string [8] the patch size is estimated from training word images.

Very recently, a method for proposing regions of interest and representing them with word string embeddings in an integrated manner has been presented [3]. The authors train a Region Proposal Network in order to predict bounding boxes. Furthermore, the predicted bounding boxes are augmented with a set of heuristically generated region proposals. A word string embedding is computed for each region. Regions are retrieved according to cosine distance with the query.

Related to word segmentation is text detection in natural scene images. These methods need to cope with large variability in the visual appearance of text. While this problem domain may seem to be less constrained compared to word segmentation, it has to be noted that the reliable detection of word boundaries in historic document images requires to correctly recognize the text in the document images first. In order to avoid recognition in our segmentation-free word spotting method, we are inspired by *extremal regions*.

Extremal regions are part of the maximally stable extremal region (MSER) blob detection method [9]. The key idea is to derive blobs based on connected components in thresholded images which are referred to as extremal regions (ER). In order to avoid the selection of a single threshold, MSERs are detected within an ER scale space. This scale space is obtained by thresholding the image at all image intensity values.

Building on the MSER approach, a method for text detection in natural scene images is presented in [10]. The method consists of different stages where character candidates are first detected, grouped into triplets and finally merged into line regions. For this purpose, ERs are extracted from color image channels. In contrast to the MSER blob detection [9], the ER stability is defined on probabilistic character class scores obtained with a boosted decision tree [10]. The final decision whether an MSER becomes a character candidate is determined with an SVM classifier.

In order to avoid the limitations of a basic connected component-based word detection, cf. e.g., [7], or patch-based frameworks, cf. e.g., [15], we propose to build ERs on top of pixel-wise text detector scores. This way, we avoid the need for classifying ERs into words and non-words which would require a word recognizer. The main advantage over a word recognizer is that the detector is applied on the entire document image and not limited to document image regions that have been heuristically selected. This way ERs model different variants for word candidates, particularly in document image regions where the detector scores are ambiguous. In order to do so, we carefully adapt the ER selection strategy. Furthermore, the integration and combination of different text detection approaches is straight forward.

To the best of our knowledge this is the first time that ERs are extracted based on detector scores. ERs have not been used in the context of segmentation-free word spotting in historic document images, before.

III. Method

Our method for segmentation-free word spotting consists of two components. In the first component word hypotheses are generated. These hypotheses are ranked with respect to similarity with the query in the second component. The overall process is depicted in Fig. 1. Text detectors produce local *detector scores* in the document image (Sec. III-A). It is an important property of our method that these scores are not binary but encode the uncertainty within the text detection process. Additionally, scores from multiple detectors can be combined. Based on the text detection scores, we obtain *word hypotheses*. This is achieved with the ER method, see Sec. III-C. The strategy for selecting word hypotheses among the ER candidates is chosen such that the most plausible word occurrences are extracted. In contrast to patch-based word spotting approaches, our method is much more robust with respect to word size variability. Finally, PHOC representations are obtained for all hypotheses through a TPP-PHOCNet [2] previously trained on segmented training images. No further adaptation is required for the segmentation-free scenario.

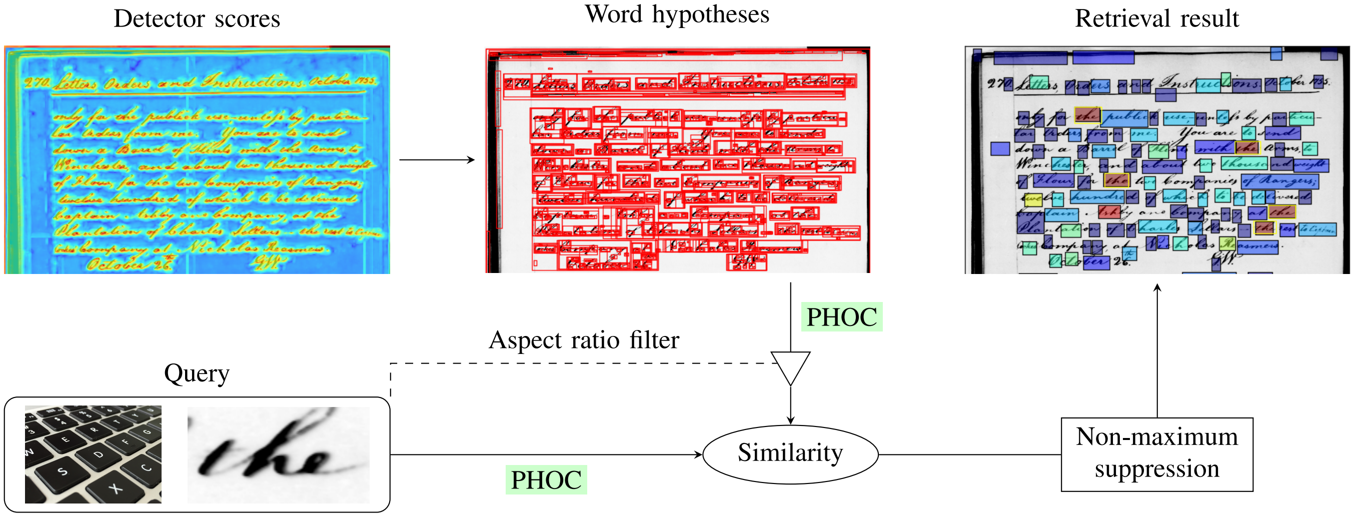

Figure 1. Word hypotheses for segmentation-free word spotting. Text detector scores indicate document image regions in a soft manner. Scores are shown with a heatmap visualization in blue to red colors. Based on the scores word hypotheses are extracted using extremal regions. Each hypothesis is represented with a PHOC and its bounding box. At query time, suitable hypotheses are obtained after aspect ratio filtering. The aspect ratio filter is depicted with a triangle. Afterwards, the similarity between the query PHOC and word hypothesis PHOCs is computed. The retrieval result is obtained after non-maximum suppression of the word hypotheses with respect to similarity with the query. The ranking is visualized with blue (dissimilar) to red (similar) colors.

For word spotting the user can provide the query either by example or by string. In order to narrow down the hypotheses to candidates that are relevant to the query, we apply a simple *aspect ratio filter* that is based on hypothesis sizes. In the query-by-string scenario the query size is estimated based on the average character width and height in the training dataset. Given the query aspect ratio a, the filter only keeps hypotheses with aspect ratios in the interval $[0.2a, 5a]$.

Afterwards, the PHOC representation of the query is computed. For textual queries the string embedding can directly be obtained. We use the TPP-PHOCNet if the query is given by example. In order to retrieve word hypotheses according to relevancy with the query, cosine similarity between the query and the hypotheses is computed. Among overlapping hypotheses the most relevant candidate is selected with *non-maximum suppression*.

A. Text Detectors

As a basis for obtaining word hypotheses in the document images we propose to use different text detector methods. We aim at finding text detectors that contain word boundary information. Different detectors can be combined by adding their predicted scores in order to improve results.

1) Dense SIFT: Text in document images is usually characterized by high contrast. Therefore, we use SIFT contrast scores in a dense grid of SIFT descriptors as text indicator. In the SIFT method the contrast scores are used for descriptor normalization [19]. The score is based on the gradient magnitudes accumulated in the descriptor cells. By using SIFT descriptors, the local text-score neighborhood can be defined. For this purpose we use descriptors consisting of 4 cells that are arranged in a single row. The cell size is 4×4 pixels. In historic document images the horizontal cell layout is useful in order to detect line boundaries. Ascenders and descenders that are touching adjacent lines obtain lower scores this way.

2) Local Region Classification (LRC): The local region classification is based on a sliding window, i.e., local regions, which is fed through a CNN. Each local region contains a classification area at the center. Based on the content of the region, the CNN classifies whether the classification area is located *inside*, *intersects* with, or is located *outside* a word bounding box. Fig. 2 visualizes the three cases. The input of the LRC detector are 64×64 pixel sized regions (black) containing 8×8 pixel sized classification areas (green), see Fig. 2. The regions are masked by a cross-shaped filter. This way, the corners are suppressed (gray) resulting in a detector for core area, ascenders, and descenders.

The CNN is based on the LeNet-5 network architecture [20]. We include two convolutional layers and one pooling layer before the LeNet-5 network and also two convolutional layers between the last pooling layer and the fully-connected layers. The output of this network is a softmax layer with three neurons representing the three classes. Text detection scores are obtained from the output neuron which is representing the class *inside*.

3) Attribute Activation Maps (AAM): The TPP-PHOCNet in its current form is able to predict a PHOC representation from a given word image. The question however is, how to locate regions in the word image that are responsible for predicting presence of attributes (i.e. characters). In order to achieve this, we draw inspiration from *Class Activation Maps (CAM)* [12]: A minor architectural change in the TPP-PHOCNet allows for extracting the region which is responsible for the attribute prediction. This is done by replacing the MLP at the end of the TPP-PHOCNet with a convolutional layer with as many filters as there are attributes. Fig. 3 visualizes

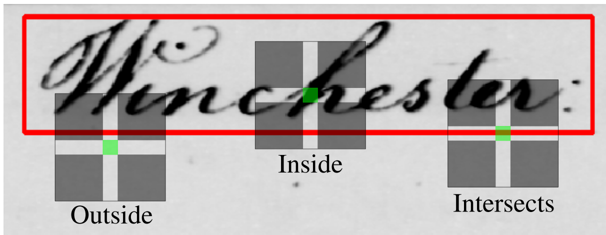

Figure 2. Exemplary input to the LRC detector representing the three classes *inside*, *intersects*, and *outside*. All parts marked with gray are ignored when processing the local regions with the CNN. Class membership is determined by the position of the crosses' intersection w.r.t. word bounding boxes.

this architectural change. Each filter in this final layer is responsible for predicting the presence or absence of one attribute. This is achieved by adding a global average pooling layer and a binary logistic loss layer to the network, similar to the CAM-model [12]. Training is still performed in an end-to-end fashion with a PHOC vector. In the spirit of [12], we term this approach *Attribute Activation Maps (AAM)* and the resulting network architecture AAM-PHOCNet.

The nice trait about the AAM-PHOCNet is that it can be trained with word images yet at test time an entire document image can be fed to the CNN. This way, we can inspect an entire document image with a single forward pass of the CNN. As each filter is trained to predict the presence of one attribute of the PHOC, the output of the AAM-PHOCNet is a 3-dimensional tensor giving a pixel-wise prediction for each attribute. For predicting character presence, we simply compute the max activation for each pixel (max over all filter responses per pixel). This produces a pixel-wise pseudo-probability. A typical per-pixel max-pooled output of the AAM-PHOCNet for a document image can be seen in Fig. 1 in the top left part (detector scores).

B. Comparing the Text Detectors

Comparing the three text detection methods from the section above, we can discriminate them with respect to their used models and required data. The SIFT-based detector does not require any training data or model at all but rather detects text in a heuristic way. In contrast, both LRC and AAM-PHOCNet require a model to be trained. The required training data, however, is quite different. The LRC requires a document image along with a set of corresponding word bounding box coordinates. A transcription of these word image regions is not required. On the other hand, the AAM-PHOCNet requires word images and their transcription but no bounding box coordinates in the document image.

C. Word Hypotheses with Extremal Regions

Based on text detector scores, we compute extremal regions (ERs). This allows for obtaining word hypotheses that are most plausible according to the text detector scores. The method is inspired by the MSER blob detector [9]. In the same spirit, we threshold *text detector scores* at a given number of thresholds. Within each thresholded image we compute connected components. Connected components computed for different thresholds are organized in a tree structure. Since text score values within a connected component are higher than the score values for surrounding pixels, these are referred to as ERs. Fig. 4 shows a tree structure of ERs for a document image section. In order to extract words, we are interested in ERs that have tree siblings. They are generated whenever local minima are found in the text scores. These minima typically represent line spaces as well as inter and intra word spaces. In order to find all relevant minima a sufficiently large number of thresholds between the minimum and maximum score must be considered.

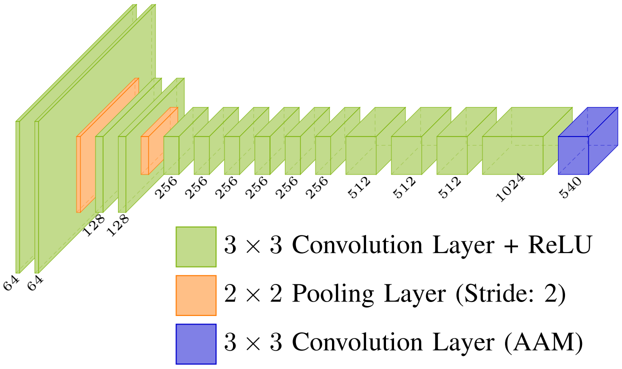
64
64
128
128
256
256
256
256
256
256
512
512
512
1024
540
🟩 3 × 3 Convolution Layer + ReLU
🟧 2 × 2 Pooling Layer (Stride: 2)
🟦 3 × 3 Convolution Layer (AAM)

Figure 3. AAM-PHOCNet architecture. In contrast to the original TPP-PHOCNet architecture, the MLP is replaced with a convolution layer featuring as many filters as there are attributes.

IV. Evaluation

For our evaluation we are describing benchmark datasets (Sec. IV-A) and evaluation protocols (Sec. IV-B). A discussion of the retrieval results achieved by our methods and a comparison with the state-of-the-art can be found in Sec. IV-C.

A. Datasets

We evaluate our method on three publicly available data sets. The first is the **George Washington (GW) data set**. It consists of 20 pages that are containing 4 860 annotated words. The pages originate from a letterbook and are quite homogeneous in their visual appearance. However, particularly for smaller words the annotation is very sloppy. As the GW data set does not have an official partitioning into training and test pages, we follow the common approach and perform a four-fold cross validation. Thus, the data set is split into batches of five consecutive documents each.

The other two data sets are referred to as **Botany** and **Konzilsprotokolle**. Both data sets were used as benchmark for the 2016 Handwritten Keyword Spotting competition [21]. For our experiments, we make use of the largest training set `Train III`. This training set contains 114 document images for Botany and 45 document images for Konzilsprotokolle. The total amount of annotated words in the training set is 16 686 for Botany and 9 102 for Konzilsprotokolle. Both data sets feature 20 test document images containing 3 318 annotated word regions for Botany and 3 891 for Konzilsprotokolle.

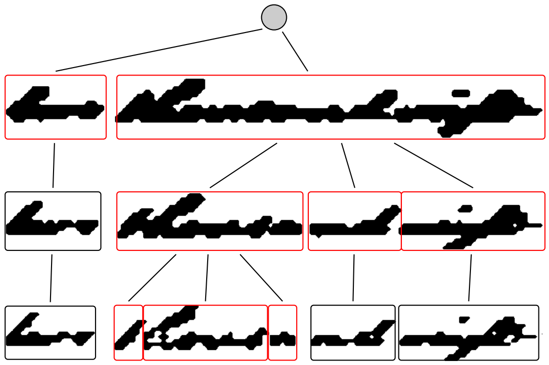

Figure 4. Exemplary word hypotheses using extremal regions. Text detector scores for a single text line are thresholded at three different values. Connected components are organized in a tree structure. The tree root is denoted by a gray circle. Hypotheses are created for tree nodes with siblings, i.e., they merge into a single parent. These are indicated with red boxes.

B. Protocol

Our protocol follows other segmentation-free word spotting protocols commonly used in the literature: Each query word image in QbE or query string in QbS is used to retrieve a list of regions from the document image collection. At training time, an annotation is given defining bounding box and transcriptions at word level. At test time, no information about word locations in the test set is given.

The retrieval lists for all queries are scored by mean average precision (mAP) as defined in [21]. In order to better assess the word hypotheses quality we also report mean recall (mR), i.e., recall averaged over all queries, and the word detection rate (DR). The DR is query-independent and is given as the relative number of word bounding boxes that have been detected by at least one word hypothesis. In our segmentation-free scenario such a detection is considered as relevant if the intersection over union of a hypothesis and a ground truth region is greater than a given overlap threshold and the retrieved region contains a word relevant to the query. The overlap threshold is 50%, unless noted otherwise, see Tab. II.

For the GW dataset, all words in the respective test split are used as queries for the QbE experiments. The QbS experiments use all unique transcriptions from the test set as queries. For Botany and Konzilsprotokolle we use the list of queries for both QbE and QbS which are defined by the benchmark.

C. Results & Discussion

The QbE results achieved with our word hypothesis methods are listed in Tab. I. The DRs for all datasets show that we obtain very accurate results. High DR is a prerequisite for high retrieval performance. Given a query, only the hypotheses can be retrieved that have been detected, beforehand.

An important result is that DR and retrieval performance can be improved when word hypothesis heights are quantized to values in $[h_{min}, h_{min}+5, \cdots h_{max}]$. These parameters are estimated such that h_{max} is the maximum word height in the training set and h_{min} is set to the typical line height in the training set. On the GW dataset h_{min} is set to 70 pixels, to 150 pixels on Konzilsprotokolle and to 120 pixels on Botany. In Tab. I these experiments are denoted with *quant*. The positive effect has mainly three reasons. First, quantization is required on GW due to the sloppy annotation of smaller words that are arbitrarily padded with white space, cf. [3]. Accurate word hypotheses will, therefore, not be considered as relevant. Second, the TPP-PHOCNet tends to favor bounding boxes that fit the text core areas. Thus, h_{min} defines a lower bound for all word hypotheses. Third, retrieval speed can be improved by suppressing similar hypotheses.

Regarding the text detectors, we evaluate the heuristic SIFT and the learned LRC and AAM methods. Further, we use linear combinations of SIFT and LRC or AAM scores, as denoted with LRC+SIFT and AAM+SIFT in Tab. I. While accurate results can be achieved with SIFT, detection and retrieval results can be improved by adding the learning-based methods. The best DRs are obtained with detectors including LRC. This is due to the explicit modeling of the visual appearance of word boundaries. Consequently, this mostly applies to retrieval performance as well. An exception can be observed on GW where the training annotations for the LRC-CNN can be considered as noisy (see above). In contrast, the AAM detector learns the visual appearance of text. The results for the AAM detectors show that the TPP-PHOCNet focusses on text core areas the most. Therefore, word hypothesis bounding boxes tend to fit closely to the words in the document.

In Tab. II we consider QbE and QbS scenarios with 50% and 25% region overlap for the segmentation-free scenario. With respect to our best performing text detector configurations, the trend in retrieval accuracy that we observed for QbE, can also be confirmed for QbS. A closer look at the results for 25% region overlap reveals that our word detections are often tighter than the original bounding box annotations. Word hypotheses that are ranked high in the retrieval list, have not been considered as relevant when using 50% region overlap.

To obtain a feasible number of word hypotheses we adjusted the number of ER-thresholds to 50 for all detectors. In our best configuration on GW (c.f. Tab. II) around 10 000 hypotheses per page were computed. After applying the aspect ratio filter approximately 5 400 regions per query and page are left for scoring. This low number of filtered hypotheses leads to an average query time of 60ms per page.

In comparison with the state-of-the-art our results compare very favourably. We outperform the previous results on Botany and Konzilsprotokolle by a large margin. On GW only the very recently presented Region Proposal CNNs [3] achieve better results. However, the authors use an additional CNN combined with brute-force hypotheses generation in order to cope with the inaccurate word annotations.

V. Conclusion

We have presented a method for segmentation-free word spotting which combines a novel ER-framework with a TPP-PHOCNet in an R-CNN framework. The ER method generates word hypotheses for which PHOCs are predicted. We proposed three different detectors in order to predict local text scores.

Table I
COMPARISON OF THE DIFFERENT TEXT DETECTION METHODS FOR THE QUERY-BY-EXAMPLE EXPERIMENTS [%]

Text detector	George Washington			Botany			Konzilsprotokolle		
	DR	mR	mAP	DR	mR	mAP	DR	mR	mAP
SIFT	88.5	73.2	64.8	85.5	75.9	66.3	93.8	89.9	86.2
SIFT$_{\text{quant}}$	93.1	88.4	80.7	88.0	77.9	68.9	96.0	91.6	87.1
LRC	92.0	81.8	77.0	90.4	80.5	71.6	89.9	81.6	76.1
LRC$_{\text{quant}}$	92.6	86.3	80.1	91.4	82.2	73.0	93.5	90.1	86.1
LRC+SIFT	92.8	82.9	78.3	91.7	82.5	73.0	96.8	92.6	88.4
LRC+SIFT$_{\text{quant}}$	93.7	88.0	81.0	93.1	84.5	**74.5**	98.5	95.2	**91.1**
AAM	51.5	35.4	31.0	73.6	63.6	53.9	91.7	77.8	70.3
AAM$_{\text{quant}}$	75.4	67.8	59.9	77.8	68.2	59.1	95.7	89.5	83.5
AAM+SIFT	88.2	68.7	62.3	87.0	76.9	67.8	96.0	88.6	84.2
AAM+SIFT$_{\text{quant}}$	93.7	89.0	**81.6**	89.0	78.6	69.4	97.7	93.4	89.6

Table II
STATE OF THE ART COMPARISON (RESULTS ARE GIVEN IN MAP [%] AT DIFFERENT OVERLAP THRESHOLDS)

Method	George Washington				Botany				Konzilsprotokolle			
	QbE		QbS		QbE		QbS		QbE		QbS	
	50%	25%	50%	25%	50%	25%	50%	25%	50%	25%	50%	25%
SIFT	64.8	71.1	70.7	76.5	66.3	75.2	68.9	79.0	86.2	91.1	84.6	91.5
SIFT$_{\text{quant}}$	80.7	90.6	82.5	89.1	68.9	76.4	72.0	80.2	87.1	94.0	87.4	92.9
LRC+SIFT	78.3	89.3	81.2	88.9	73.0	79.9	76.2	83.6	88.4	94.9	86.6	95.6
LRC+SIFT$_{\text{quant}}$	81.0	92.0	83.6	90.5	**74.5**	**80.4**	**78.8**	**85.3**	**91.1**	95.6	**89.9**	95.3
AAM+SIFT	62.3	69.0	70.0	76.1	67.8	75.4	71.0	80.1	84.2	94.9	81.9	95.2
AAM+SIFT$_{\text{quant}}$	81.6	92.0	84.6	90.6	69.4	75.9	74.0	80.3	89.6	**96.2**	88.9	**96.0**
BoF-HMM [16]	–	–	76.5	80.1	–	–	–	–	–	–	–	–
Ctrl-F-Net [3]	**90.9**	**97.0**	**91.0**	**95.2**	–	–	–	–	–	–	–	–
TAU [21]	–	–	–	–	37.48	–	–	–	61.78	–	–	–
Attribute-SVMs+RR [8]	–	–	73.7	–	–	–	–	–	–	–	–	–

This way, we avoid using a patch-based framework as well generating large amounts of region hypotheses blindly. In the experimental evaluation we achieve results that are competitive with the state-of-the-art.

REFERENCES

[1] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *TPAMI*, vol. 36, no. 12, pp. 2552–2566, 2014.

[2] S. Sudholt and G. Fink, "Evaluating word string embeddings and loss functions for CNN-based word spotting," in *Proc. of the ICDAR*, Kyoto, Japan, 2017, to appear.

[3] T. Wilkinson, J. Lindström, and A. Brun, "Neural Ctrl-F: Segmentation-free query-by-string word spotting in handwritten manuscript collections," *ArXiv e-prints*, Mar. 2017.

[4] R. Manmatha and J. L. Rothfeder, "A scale space approach for automatically segmenting words from historical handwritten documents," *TPAMI*, vol. 27, no. 8, pp. 1212–1225, Aug 2005.

[5] T. Wilkinson and A. Brun, "A novel word segmentation method based on object detection and deep learning," in *Int. Symposium of Advances in Visual Computing*, 2015, pp. 231–240.

[6] J. Rodríguez-Serrano and F. Perronnin, "Handwritten word-spotting using hidden Markov models and universal vocabularies," *Pattern Recognition*, vol. 42, no. 9, 2009.

[7] A. Kovalchuk, L. Wolf, and N. Dershowitz, "A simple and fast word spotting method," in *Proc. of the ICFHR*, Sept 2014, pp. 3–8.

[8] S. K. Ghosh and E. Valveny, "Query by string word spotting based on character bi-gram indexing," in *Proc. of the ICDAR*, Aug 2015, pp. 881–885.

[9] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761 – 767, 2004.

[10] L. Neumann and J. Matas, "Real-time lexicon-free scene text localization and recognition," *TPAMI*, vol. 38, no. 9, pp. 1872–1885, Sept 2016.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *TPAMI*, vol. 38, no. 1, pp. 142–158, 2016.

[12] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. of the Conf. on CVPR*, 2016, pp. 2921–2929.

[13] Y. Leydier, A. Ouji, F. LeBourgeois, and H. Emptoz, "Towards an omnilingual word retrieval system for ancient manuscripts," *Pattern Recognition*, vol. 42, no. 9, 2009.

[14] B. Gatos and I. Pratikakis, "Segmentation-free word spotting in historical printed documents," in *Proc. of the ICDAR*, 2009.

[15] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Segmentation-free word spotting with exemplar SVMs," *Pattern Recognition*, vol. 47, no. 12, pp. 3967 – 3978, 2014.

[16] L. Rothacker and G. A. Fink, "Segmentation-free query-by-string word spotting with bag-of-features HMMs," in *Proc. of the ICDAR*, Aug 2015, pp. 661–665.

[17] M. Rusiñol, D. Aldavert, R. Toledo, and J. Lladós, "Efficient segmentation-free keyword spotting in historical document collections," *Pattern Recognition*, vol. 48, no. 2, pp. 545 – 555, 2015.

[18] S. K. Ghosh and E. Valveny, "A sliding window framework for word spotting based on word attributes," in *Pattern Recognition and Image Analysis*, 2015, vol. 9117, pp. 652–661.

[19] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. Journal of Computer Vision*, vol. 60, 2004.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[21] I. Pratikakis, K. Zagoris, B. Gatos, J. Puigcerver, A. Toselli, and E. Vidal, "ICFHR2016 Handwritten keyword spotting competition (H-KWS 2016)," in *Proc. of the ICFHR*, 2016, pp. 613–618.