Grouping Historical Postcards Using Query-by-Example Word Spotting

Gernot A. Fink, Leonard Rothacker, René Grzeszick
Department of Computer Science
TU Dortmund University
Dortmund, Germany
{Gernot.Fink|Leonard.Rothacker|Rene.Grzeszick}@udo.edu

***Abstract*—Handwritten historical documents pose extremely challenging problems for automatic analysis. This is due to the high variability observed in handwritten script, the use of writing styles and script types unknown today, the frequently lacking orthographic standardization, and the degradation of the respective documents. Therefore, it is currently out of question to develop general purpose handwriting recognition systems for historical document collections. It is, however, possible to search relatively homogeneous document collections using word spotting techniques. In this paper we consider the analysis of a challenging collection of postcards from the period of World War I delivered by the German military postal service. More specifically, we consider the automatic grouping of mail pieces by spotting potentially identical addressees. As the annotation of such documents is extremely challenging even for trained experts, a manually developed ground truth annotation will, in general, not be available. Furthermore, a reliable segmentation on word level will hardly be possible. With our segmentation-free query-by-example word spotting method we investigate modifications addressing the better generalization to a multi-writer scenario and its application to degraded documents. Promising results could be achieved in this highly challenging scenario.**

***Keywords*-word spotting; historical documents; handwriting recognition;**

I. Introduction

The automatic analysis and recognition of handwritten documents is still an open research problem. Over the last decades huge progress has been made by applying statistical methods and machine learning techniques which are capable of learning complex models of handwriting appearance from example data (cf. e.g. [1]). In combination with statistical models for the language fragment investigated, these techniques are capable of achieving impressive transcription accuracies on isogeneous document collections (cf. e.g. [2]).

Historical handwritten documents pose additional challenges to the modeling of the high variability observed in handwriting. Due to the aging and possibly improper handling of the documents over time, the document itself might be degraded. Additionally, the documents will typically be written in writing styles or even script types that are no longer in use today. Therefore, it will be impossible to apply a general purpose handwriting recognizer trained on contemporary material for the transcription of historical handwritten documents. Moreover, language model restrictions might be hard to exploit as historical texts usually do not adhere to rules of orthographic standardization in the same way as modern texts do. In consequence, collections of handwritten historical documents will show a high degree of "individuality" and it will be extremely problematic to collect large amounts of annotated training material, which shows the same overall appearance of the script, in order to estimate a powerful statistical recognition model.

In order to be able to at least partly analyse document collections without the need to prepare large amounts of annotated material of the same type in advance, ideas from the field of image retrieval led to the development of query-by-example word spotting techniques (cf. e.g. [3], [4], [5]). The general idea is that the user selects a query word-image from a document in a certain collection in an interactive process. Based on this query image the document collection is then searched for image patches with similar appearance and a ranked list of retrieval results is returned. When assuming that the appearance of a query word will vary only little within the document collection considered, quite impressive results can be achieved with such techniques. As query-by-example word spotting will require no prior training of a model, the technique can also be applied for the analysis of previously unexplored document collections for which little or no annotations are available yet.

In this paper we apply our state-of-the-art word spotting method [6] for the analysis of a challenging collection of postcards from the period of World War I delivered by the German military postal service. More specifically, we consider the automatic grouping of mail pieces according to the respective addressees by spotting potentially identical addressees starting from the name part of an address found in a query document. In contrast to classical applications of query-by-example word spotting, we address the following challenges here: First, the number of writers writing to the same person is unknown a priori and we, therefore, have to consider a multi-writer scenario. Second, the evaluation of the method cannot rely on a completely available ground truth transcription due to the effort required, such that evaluation metrics have to be approximated appropriately.

II. Related Work

Due to the high variability observed in handwriting and the frequently rather poor quality of historical documents, today's automatic transcription techniques mostly fail on this class of documents and transcriptions have to be generated manually. Word spotting was first proposed in [7] in order to reduce the manual annotation effort by automatically clustering similar word images. In this and subsequent works (cf. e.g. [8]) it was assumed that documents could be segmented into individual word images. The matching between these was then performed by applying dynamic time warping on image profile features. Later, the basic idea was enhanced by developing more powerful feature representations based on local image descriptors such as SIFT and HOG and by applying the Bag-of-Features principle for representing the appearance of query word images [3], [4], [5]. In [4], a query representation similar to the spatial pyramid proposed by [9] and a patch-based decoding of the query model was introduced such that the segmentation of documents into word images prior to word spotting could be avoided. Recently, we proposed a combination of the Bag-of-Features representation of word images with hidden Markov models [6], the so-called Bag-of-Features hidden Markov models (BoF-HMMs). This approach can be seen as a generalization of [4] introducing the flexibility of a probabilistic sequence model.

The main advantage of all the word spotting methods outlined above is that they do not require any supervised training of the query models, which means that no tedious annotation process is necessary. The query models can be built on the fly and historical document collections can be searched in a query-by-example approach. This procedure has the clear limitation that the document databases considered need to exhibit only very low variation in the appearance of the handwriting. Therefore, query-by-example methods are especially well suited for single-writer tasks and document collections that were produced in highly similar writing styles by a small number of writers only.

The most important disadvantage of the query-by-example principle is, however, that only queries can be used for which an example can be selected from a given document. Using arbitrary text queries is not possible as there is no model of the relation between character classes and their appearance. So-called query-by-string word spotting methods try to create such models, e.g., by composing query models from pre-trained character models [10] or by estimating a mapping between visual and textual representations [11]. Consequently, all query-by-string word spotting approaches proposed so far require substantial amounts of manual annotations of data before being applicable and, therefore, are not suitable for the exploration of yet undeveloped document collections, as considered in this research, for which little or no annotations are available.

III. German Feldpost in World War I

The First World War is commonly considered to be the first industrialized war in history. In addition to a massive use of technical resources both at the front-line and in the support of the troops by civilians at the so-called home front, it also brought a substantial development in the military postal services. These were supplied – mostly free of charge – for communication between soldiers and their relatives and friends at home. Almost 100 years ago this so-called *feldpost* could be considered the equivalent of today's social media and was frequently used on a daily basis. On the German side, in total approximately 28.7 billion mail pieces were delivered from the front to the homeland and vice versa. This huge volume of mail – approximately 16 million deliveries per day – includes parcels, feldpost letters, and approximately 25% of postcards ([12, p. 29]).

Though feldpost from World War I can be considered as a valuable resource for historical, cultural, and linguistic research, it still is largely unexplored and the collections available are limited to a few thousand mail pieces each (cf. e.g. [13]). What sets these documents apart from other historical manuscripts is that they can be considered as records of historical every-day communication. Consequently, individual items will be of relatively limited interest and the analysis has to consider larger volumes of such documents. This is where automatic analysis comes into play as manual transcriptions of substantial amounts of such historical mass-sources are not feasible.

IV. Word Spotting for Grouping Historical Postcards

A. Task Definition

Figure 1 shows examples of feldpost postcards from the collection [14] considered in this paper. When analyzing such a document collection, a first high-level task is to group the documents according to the people involved in the communication process. Even though the writing of addresses varies largely within these documents, the name of the addressee can usually be identified clearly, though the actual transcription might be non-trivial. Interpreting a given word image of a person's last name as a query, word spotting techniques can be applied in order to search for relevant matches within the document collection. The query document and all correctly identified matches returned by the search constitute a group of documents.

As feldpost communication took place between relatives and friends, the number of people writing to a person will be limited and therefore also the variation in the appearance of the name written will be limited making a query-by-example approach feasible. However, there is a special peculiarity in writing observed in these documents from the beginning of the 20*th* century, which makes things more complicated. As can be seen from the examples in Fig. 1, the overall

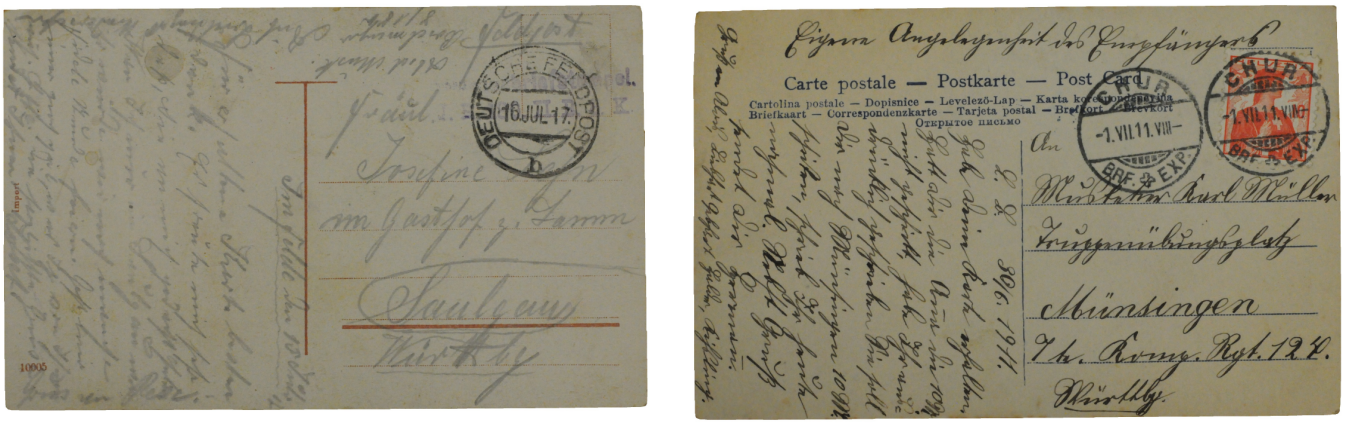

Figure 1. Example of German feldpost postcards from World War I (from [14], picture side not shown): On the left, the recipient's name "Josefine Lehn" is mostly written in Roman script but with a Kurrent "h". On the right, the address of "Karl Müller" is mostly written in Kurrent.

script type used at that time was German Kurrent. Addresses and names of people, however, were frequently written in the "modern" Roman script with mixups occurring due to confusions between letter shapes.

B. General Methodology

In order to automatically retrieve postcards addressed to the same person, we make the following assumptions: First, we assume that query-by-example word spotting is the only realistic tool that can be employed as methods for query-by-string or even transcription would require the annotation of substantial amounts of material from the same document collection. Second, we assume that segmentation-based approaches will in general fail on the type of documents considered here (cf. Fig. 1) as even for the address part a reliable word segmentation will not be feasible.

Therefore, we apply our state-of-the-art segmentation-free query-by-example word spotting approach [6]. In order to match documents based on the last names found in the address part, we perform the processing steps described in the following sub-sections.

C. Preprocessing

First, document images are preprocessed in order to improve the overall contrast between the script and the document background, which is frequently quite low due to fading of the script or document degradations. We apply histogram equalization to the intensity channel in an YCrCb color space. Then, a 9×9 median filter is applied for reducing the background noise. An example of the improvement that is achieved is shown in Fig. 2.

D. Local Image Features

Similarly to [3], [4], [5] we compute SIFT descriptors centered on image positions arranged in a regular grid with fixed spacing. Subsequently, all SIFT descriptors obtained for a document collection are subject to a clustering process

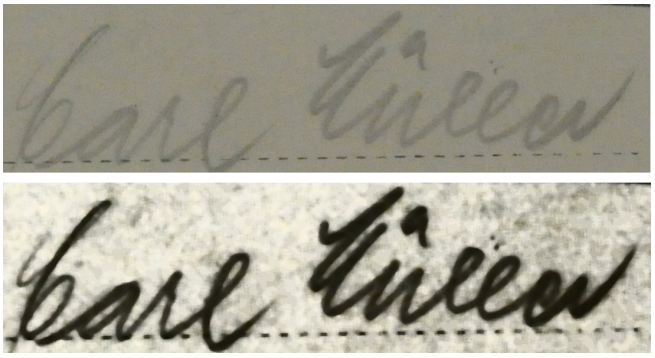

Figure 2. Example of preprocessing. Top: original image ("Karl Müller") with low contrast. Bottom: improvements after histogram equalization and median filtering.

by applying the generalized Lloyd algorithm. The resulting codebook forms the visual vocabulary used for quantizing the image descriptors.

E. Query Modeling

Query words, for which the respective word images have been extracted manually from a document image, are modeled using Bag-of-Features HMMs [15]. Following the modeling strategy used for HMM-based handwritten text recognition (cf. e.g. [1]), a sliding window is moved over the query word-image from left to right. For every analysis window obtained, a histogram of the quantized image descriptors contained within that window is computed. The sequence of these term vectors then forms the input data to a BoF-HMM. This model is created with linear topology and a number of states proportional to the length of the term-vector sequence obtained for the query. The model parameters are then estimated in a procedure similar to Viterbi training starting with uniformly initialized parameters. Please note that the estimation of a single BoF-HMM works on a single word image that has been manually provided as a query.

F. Segmentation-Free Matching

For matching queries with potential target regions within a document in a segmentation-free manner, a patch-based approach is applied. Patches are densely sampled on the document level and for each patch a sequence of term vectors is extracted from the dense grid of quantized image descriptors. The patch size is specific to the size of the query word image. Given a BoF-HMM query model, the likelihood of generating a patch's term vector sequence can be decoded with the Viterbi algorithm. This probabilistic score indicates the similarity to the query. Patches with locally maximal scores are sorted by similarity and returned as retrieval result. The application of a patched-based framework and Bag-of-Features representations to segmentation-free word spotting was first proposed in [4]. More details about BoF-HMMs and their application to segmentation-free word spotting can be found in [15], [6].

V. Evaluation

A. Dataset of Historical Postcards

The dataset of German feldpost postcards considered in this work is part of a private collection of postcards from World War I [14]. The collection focuses on mail items related to the western war zone, i.e., postcards sent to and from the German front-lines in Belgium, and northern France. The collection comprises mostly postcards from bequests such that social relations of persons can be observed in the data.

In total 1,346 postcards from nine different acquisition campaigns have already been digitized. All postcards were photographed at approximately 600 dpi with a total resolution of 4288×2484 pixels. The photographs were taken in front of a red background that is removed by color space thresholding. The postcards themselves are approximately 3200×2100 pixels in size.

B. Evaluation Tasks

In this work, two sub-sets of this collection are considered, which consist of 100 and 460 postcard images, respectively.

For the first set of 100 postcards we manually prepared the ground truth annotation for all last names found in the address parts of the documents. These annotations are defined by bounding boxes on the word level together with the transcription. As with respect to the name-spotting task the complete annotation is available for this set, we refer to it as the *Closed Set* in the following [1]. For evaluation purposes it can be used in much the same way as the segmentation-free word spotting benchmark defined on the George Washington dataset [4] where every single item of the ground truth annotation is considered as a query and the evaluation results obtained are averaged over all queries.

[1]The closed set benchmark (document images and ground truth annotation) will be made available for research purposes on request.

For the second set of 460 postcards ground-truth information is not available. Therefore, we refer to it as the *Open Set*. Though evaluation metrics can only be partly estimated for such a data set, it represents a quite realistic real-word situation where scholars just start to develop a collection of historical documents. In the initial phase of such a project, annotations will be missing completely and the quality of automatic procedures used to support the dataset development process can only be judged subjectively. We simulate such a situation on the *Open Set* by manually choosing a set of 10 queries. For each query the 10 best candidate matches are computed, which is a typical number of retrieval results returned by internet search engines.

C. Evaluation Metrics

The input to a word spotting system is a query word image. The final output is a list of image regions that are ranked according to their similarity with respect to that query. For a potential user an ideal word spotting system should present a list that meets two criteria. All relevant results should be listed first. Secondly, the list should contain all relevant items that are present in the data collection. These two conditions compete against each other. The chance of having a well sorted list is higher when the list is short. On the other hand shorter lists tend to contain fewer relevant results.

In terms of evaluation measures the first criterion is described by average precision and the second criterion by recall. When evaluating a word spotting system for a set of queries, the performance is described by mean average precision (mAP) and mean recall (mR).

For the *Closed Set* full ground truth is available. For each of the 100 queries we extract the top ten patches per postcard and compute average precision and recall on the resulting ranked lists. The evaluation protocol is oriented at and comparable to the evaluation of the George Washington benchmark (cf. [4], [6]). A patch is considered as relevant if it overlaps with a corresponding ground truth annotation by more than 10%. This low overlap threshold also allows detections for varying word sizes. In the evaluation it favors mean recall and disfavors mean average precision.

In order to evaluate the *Open Set*, only approximate measures can be given as annotations are only available for the manually selected queries. On each postcard we extract the patch that is most similar to the query. The resulting list of the top ten postcards per query is manually evaluated with respect to relevance. The mean average precision gives a rough estimate of the performance. As the total number of occurrences of each query in the Open Set is generally unknown, the mean recall cannot be estimated.

D. Results

Results for spotting family names on historical postcards are given for the *Closed Set* and the *Open Set*. As query-by-example word spotting models have to be estimated from

Table I
WORD SPOTTING PERFORMANCE

Task	Queries / Pages	Preprocessing	Descriptor	Visual Vocabulary	Mean Average Precision	Mean Recall
Closed Set	100 / 100 (incl. query resp.)	none	60x60	2048	26.6 %	61.3 %
Closed Set	100 / 100 (incl. query resp.)	hist equ	60x60	2048	26.0 %	61.9 %
Closed Set	100 / 100 (incl. query resp.)	hist equ + median	60x60	2048	30.1 %	75.7 %
Closed Set	100 / 100 (incl. query resp.)	hist equ + median	40x40	2048	27.2 %	75.9 %
Closed Set	100 / 100 (incl. query resp.)	hist equ + median	60x60	2048	30.1 %	75.7 %
Closed Set	100 / 100 (incl. query resp.)	hist equ + median	80x80	2048	29.3 %	76.5 %
Closed Set	100 / 100 (incl. query resp.)	hist equ + median	60x60	1024	28.2 %	76.0 %
Closed Set	100 / 100 (incl. query resp.)	hist equ + median	60x60	2048	30.1 %	75.7 %
Closed Set	100 / 100 (incl. query resp.)	hist equ + median	60x60	4096	29.7 %	77.5 %
Open Set	10 / 460 (incl. query resp.)	hist equ + median	60x60	2048	81.8 %	-
Open Set	10 / 460 (without query resp.)	hist equ + median	60x60	2048	22.9 %	-

a single sample, we investigate parameters that are suitable for generalizing the model to handle more variability. In the given scenario addressees are written by multiple writers leading to different writing styles, size, etc. Additionally, we report the effect of the considered preprocessing methods.

The parameter optimization is performed on the *Closed Set*. With a fully available ground truth the parameter space can be explored systematically. Tests on the *Open Set* are finally performed with the best configuration found in the prior validation. Table I shows the results for both tasks. In compliance with the benchmark for segmentation-free word spotting on the George Washington benchmark (cf. [4]), retrieval lists can also contain the query itself. This biases the results positively, but makes hardly any difference for long retrieval lists. However, for shorter lists like in our *Open Set* scenario the difference is substantial. Over 80 % mean average precision is completely unrealistic in comparison to the 30 % that we can report as best result in the *Closed Set* validation. For that reason we manually identified the query responses in the retrieval lists and excluded them for computing *Open Set* performance measures. In the segmentation-free framework the query is mostly found as the top hit. This is not always guaranteed. Due to the dense sampling of patches (cf. Sec. IV-F), the exact term vector sequence used to estimate the query model is not always found. Fig. 3 shows exemplary retrieval results for three out of ten queries in the *Open Set*. In these cases the query itself is retrieved as the top hit.

As reported for many recognition tasks dealing with degraded historical documents, preprocessing helps to reduce unwanted variabilities that are caused by document degradations. In our scenario we frequently find postcards with low contrast (cf. Fig. 2). The results in Table I show that the sole application of histogram equalization does not improve word spotting performance notably. Although the contrast of the pen stroke is greatly improved, the noisy postcard background is overly emphasized. We handle this problem by applying an edge preserving median filter. As Table I shows, this leads to substantially better results for both mean average precision and mean recall.

The size of the descriptor and the size of the visual vocabulary are both important for the model's ability to generalize from a single sample of the query word to other occurrences of the same word written in different styles. The larger the descriptor, the more specific the overall representation. If a single descriptor captures whole groups of characters, it describes those characters' shapes and how they are connected. On the other hand, if descriptors capture only parts of characters the representation gets less specific to whole character shapes. Intuitively, for single-writer scenarios better performance can be achieved with larger descriptors. Here, we found a trade off between specificity and generality for our multi-writer query-by-example scenario with 60x60 pixels descriptors. The situation is very similar for the size of the visual vocabulary. Its items are the visual representatives that the Bag-of-Features representation is built upon. Quantizing descriptors with respect to these representatives generalizes the representation to the visual features that can typically be observed in the given dataset (cf. Sec. IV-D). For fewer representatives the level of generalization increases as it decreases with higher numbers of visual representatives. The trade-off for the vocabulary size is found at 2,048. Here, we give the mean average precision more emphasis than the mean recall. For users a good precision in the top ranks of the retrieval list is typically more important than its completeness.

Finally, we give the results for the *Open Set* task. Although, we can only give a rough estimate for the mean average precision, the result of 22.9 % seems realistic in comparison to the word spotting performance on the validation. The mean recall cannot be estimated due to missing ground truth. However, for two queries the number of occurrences in the data collection is known. Their recall scores of 36 % and 30 % give a rough idea of the performance that can be expected for a 10 element retrieval list in the given scenario. Fig. 3 visualizes these two and a third query in order to show

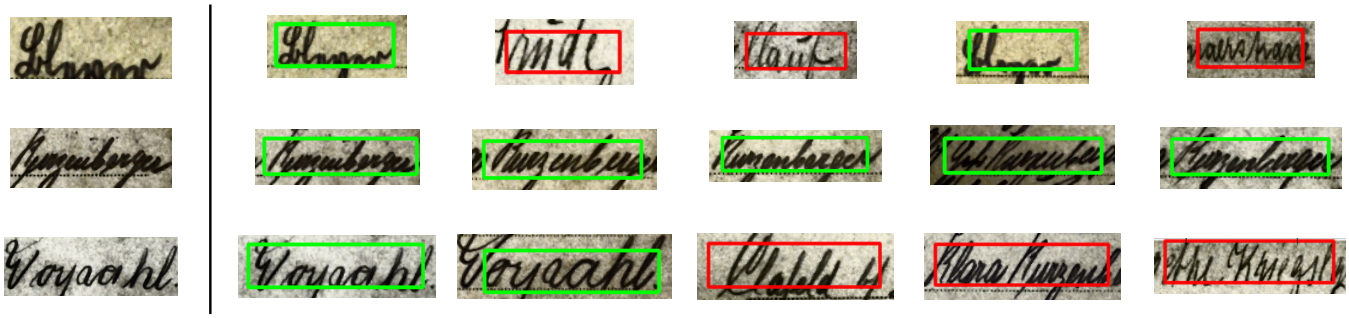

Figure 3. Exemplary queries (left) and their top five retrieval responses (right) from the *Open Set*. Positive and negative hits are visualized with green and red boxes respectively. Note that the query itself will usually be retrieved as the top hit in our segmentation-free framework.

some qualitative results. The choice of the queries should demonstrate cases of both success and failure.

VI. Conclusion

In this paper we presented a method for grouping historical postcards based on addressees using segmentation-free query-by-example word spotting. The task is extremely difficult since the script is diverse and the available ground truth is very limited. It is neither possible to segment the data nor to annotate a sufficient amount of samples to train a more sophisticated recognizer. Even for trained experts transcribing the text is very tedious work. Therefore, the presented state-of-the-art word spotting approach is a feasible alternative. Optimizing the method for better generalization to different writing styles showed promising results in a closed setup as well as in a large open setup without full ground truth information.

Acknowledgment

The authors would like to thank Dr. Britta Bley, Dortmund, Germany for providing the collection of historical postcards and for her consultations regarding historical aspects of the document type considered here.

References

[1] T. Plötz and G. A. Fink, "Markov Models for Offline Handwriting Recognition: A Survey," *Int. Journal on Document Analysis and Recognition*, vol. 12, no. 4, pp. 269–298, 2009.

[2] M. Kozielski, P. Doetsch, and H. Ney, "Improvements in RWTH's system for off-line handwriting recognition," in *Proc. Int. Conf. on Document Analysis and Recognition*, Washington, DC, USA, 2013.

[3] J. A. Rodríguez-Serrano and F. Perronnin, "Handwritten word-spotting using hidden Markov models and universal vocabularies," *Pattern Recognition*, vol. 42, no. 9, pp. 2106–2116, September 2009.

[4] M. Rusiñol, D. Aldavert, R. Toledo, and J. Llados, "Browsing heterogeneous document collections by a segmentation-free word spotting method," in *Proc. Int. Conf. on Document Analysis and Recognition*, Bejing, China, 2011, pp. 63 –67.

[5] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Efficient exemplar word spotting," in *Proc. British Machine Vision Conference*, 2012, pp. 67.1–67.11.

[6] L. Rothacker, M. Rusinol, and G. A. Fink, "Bag-of-features HMMs for segmentation-free word spotting in handwritten documents," in *Proc. Int. Conf. on Document Analysis and Recognition*, Washington DC, USA, 2013.

[7] R. Manmatha, C. Han, and E. M. Riseman, "Word spotting: a new approach to indexing handwriting," in *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, Jun 1996, pp. 631–637.

[8] T. M. Rath and R. Manmatha, "Word image matching using dynamic time warping," in *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, vol. 2, June 2003, pp. 521–527.

[9] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2169–2178.

[10] A. Fischer, A. Keller, V. Frinken, and H. Bunke, "Lexicon-free handwritten word spotting using character hmms," *Pattern Recogn. Lett.*, vol. 33, no. 7, pp. 934–942, May 2012.

[11] D. Aldavert, M. Rusinol, R. Toledo, and J. Llados, "Integrating visual and textual cues for query-by-string word spotting," in *Proc. Int. Conf. on Document Analysis and Recognition*, 2013, pp. 511–515.

[12] C. Brocks, *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)*. Essen: Klartext-Verlag, 2008, (in German).

[13] "Europeana 1914–1918 — untold stories & official histories of WW1," http://www.europeana1914-1918.eu.

[14] B. Bley, "Feldpostkarten im 1. Weltkrieg (Feldpost Postcards of World War I)," Private Collection, Dortmund, Germany.

[15] L. Rothacker, S. Vajda, and G. A. Fink, "Bag-of-features representations for offline handwriting recognition applied to arabic script," in *Proc. Int. Conf. on Frontiers in Handwriting Recognition*, Bari, Italy, 2012.