Word Spotting in Historical Document Collections with Online-Handwritten Queries

Christian Wieprecht, Leonard Rothacker, Gernot A. Fink
Department of Computer Science
TU Dortmund University
Dortmund, Germany
Email: {firstname.lastname}@udo.edu

***Abstract*—Pen-based systems are becoming more and more important due to the growing availability of touch sensitive devices in various forms and sizes. Their interfaces offer the possibility to directly interact with a system by natural handwriting. In contrast to other input modalities it is not required to switch to special modes, like software-keyboards. In this paper we propose a new method for querying digital archives of historical documents. Word images are retrieved with respect to search terms that users write on a pen-based system by hand. The captured trajectory is used as a query which we call query-by-online-trajectory word spotting. By using attribute embeddings for both online-trajectory and visual features, word images are retrieved based on their distance to the query in a common subspace. The system is therefore robust, as no explicit transcription for queries or word images is required. We evaluate our approach for writer-dependent as well as writer-independent scenarios, where we present highly accurate retrieval results in the former and compelling retrieval results in the latter case. Our performance is very competitive in comparison to related methods from the literature.**

***Index Terms*—word spotting; pen-based systems; online handwriting representations; common subspaces**

I. Introduction

Word spotting is the task of retrieving words from digital document collections with respect to a query given by the user [1]. This makes these archives searchable without transcribing the documents first. Especially for historical data it is hard to obtain a full transcription automatically. This is mainly due to the lack of large annotated training corpora. However, such a transcription would be required for rapidly exploring these collections. Word spotting systems are also more robust to recognition errors. Their output is a list of word images ranked according to similarity to the query. Retrieval errors mainly effect the ranking of the list which can be dealt with quite easily by the users. In contrast, errors in automatic transcriptions usually make the recognition result useless [2].

In this paper we propose a new pen-based interface for word spotting systems. In the past years, touch sensitive devices have become very popular. Intuitive user interfaces have been developed for smartphones and tablets, smartboards can be found in classrooms and offices, and TV-sized touch panels have become popular for interacting with visitors of museums and exhibitions. Although the standard interaction is often based on touch gestures, especially smartboards and tablets specifically offer pen-based interfaces. Currently, there are

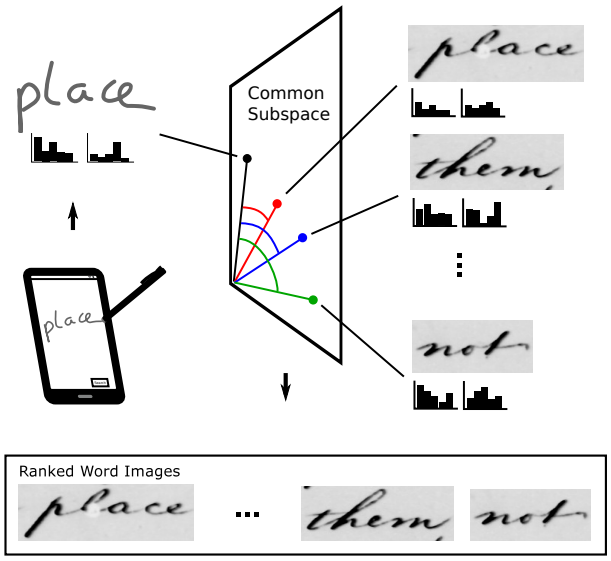

Fig. 1. Query-by-online-trajectory word spotting. Visual word image features as well as the query's online-handwritten trajectory feature representation are projected into a common subspace. Word images (from [4]) are retrieved based on their distance to the query which is indicated by differently colored angles.

mainly two ways of processing the pen-based input. The pen trajectory can either be automatically transcribed using an online-handwriting recognition system or rendered and embedded as an image with respect to the context of the application, e.g., as an annotation within a document. The latter option does not include any automatic interpretation and is, therefore, only intended for users and not for further interpretation by other systems. In contrast, the former option performs a complete interpretation resulting in a machine-based transcription that can principally be used in any subsequent system. While this is certainly desirable, it is hard to achieve in practice. Online-handwriting recognizers require huge amounts of training material, large lexica and specifically adapted methods in order to cope with the great variability found in unconstrained human handwriting [3].

Here, we propose an in-between solution supporting the interpretation of online-handwritten trajectories for word spotting. This avoids problems of full transcription recognizers while users are still aided with the possibility to search for

words, which is one of the most important functionalities when working with large document archives. For this purpose we propose to use separate embeddings for projecting both visual features, representing word images, and online-handwriting features, representing queries, into a common subspace that directly allows distance-based retrieval. The full process is depicted in Figure 1. In order to embed features from different domains, we build on top of recent advances in text recognition and query-by-string word spotting. In [5], [6] methods were presented that embed visual features and textual features into a common subspace. While their textual representations are derived from ground truth information, we have to cope with the variability found in unconstrained online-handwriting in order to create a writer independent user interface. Furthermore, we evaluate a different common subspace method [7] that has recently been proposed for query-by-speech word spotting.

The rest of this paper is organized as follows. In Section II we give a brief overview of related word spotting methods and a more detailed discussion of approaches allowing cross-domain retrieval, i.e., [7], [8] and [6]. Section III describes our proposed method for query-by-online-trajectory word spotting. This particularly includes a new feature representation for online-handwritten words and the query-by-online-trajectory attribute embedding pipeline. We present the results of our evaluation in Section IV and give a conclusion in Section V.

II. Related Work

From the user's perspective a word spotting system's interface is a very important component for its efficient application when exploring document collections. While this paper proposes a method for using online-handwritten keyword queries, there are two popular ways of defining word spotting queries in the related literature. The first is referred to as query-by-example. An exemplary instance of the query word is provided by the user and word image retrieval is then based on visual similarity between the query word image and local regions in the document images, e.g., [9]–[11]. On the downside, this is quite inconvenient for the user because the query word image has to be located first, a task that can be quite time-consuming for infrequent words (also compare [7], [8]). On the upside, no annotated training material is required. The second approach is referred to as query-by-string. This is much more convenient for the user because the query can simply be entered as a textual representation with a keyboard. Depending on the scenario there are different approaches to query-by-string word spotting. If the variability in the script's visual appearance is very low, template-based methods are sufficient. For example, in [12] query word images are generated automatically by concatenating character templates that are given manually. Afterwards, query-by-example techniques can be applied. However, if the variability of the script's visual appearance is high, textual models must be estimated with large amounts of annotated training material. Suitable methods are, for example, based on full transcription recognizers. In [2], [13] text line images are ranked according to their probability of containing the query word. Fischer et al. [13] present an HMM-based system whereas Frinken et al. [2] use recurrent neural networks.

In the past years, word spotting methods emerged that are inspired by techniques from natural language processing and computer vision. This is mainly due to the adaption of the Bag-of-Words model and higher order methods built on top of them. Rusiñol et al. [10] indexed Spatial Pyramid Bag-of-Features representations in a topic space with Latent Semantic Indexing (LSI) allowing for the rapid retrieval of document image regions in a segmentation-free query-by-example word spotting framework. They extended their idea to query-by-string word spotting in [8] by modeling correlations between textual and visual features in this common subspace. Another common subspace method for word spotting was presented by Almazán et al. [6]. They use an attribute embedding which they base upon their PHOC (Pyramidal Histogram Of Characters) representations. Textual features can directly be encoded in PHOC space whereas visual features are embedded with PHOC-specific Attribute-SVMs (Support Vector Machines). Visual features are represented by Fisher Vectors which are enriched with spatial information.

In the remainder of this section these two methods will be discussed in more detail as they are both suitable for cross-domain word spotting, i.e., queries can originate from any domain including others than the visual or textual.

Aldavert et al. [8] proposed an LSI-based query-by-string word spotting method. For a set of transcribed word images visual and textual feature representations are calculated during training. For each instance both feature representations are concatenated in order to form a single descriptor. Using Singular Value Decomposition a vector subspace is calculated in which the correlations between visual and textual features are encoded. At retrieval time no textual information (i.e. transcription) is available for word images from the document collection. For the textual query, in contrast, no visual feature information is present. However, feature representations of the query and all word images from the document collection can be projected into the same topic subspace by setting the respectively missing representation to zero. Thus, retrieval becomes a nearest neighbor problem in the common subspace.

Almazán et al. [6] proposed a query-by-string word spotting method based on a common PHOC subspace. A PHOC vector represents the distribution of characters in a word by indicating their positional occurrence in a binary manner. For this purpose the word is divided into different regions in a pyramidal structure, e.g., split in two regions in one and in three regions in the next level. Binary flags for all characters indicate their presence within each region. This is conceptually corresponding to Spatial Pyramids that are very common for encoding spatial information in images. Each dimension of the PHOC vector defines an attribute for the attribute embedding. During training a PHOC vector is determined for the transcription of each word image. For each of those attributes the visual feature representations of all word images are then grouped into a positive and a negative set, depending on whether the corresponding attribute in the PHOC vector is active. With these

sets individual SVMs are trained for each attribute which are therefore called Attribute-SVMs. For projecting word images' visual feature representations into PHOC space at retrieval time, the visual feature vector is scored by each Attribute-SVM. These scores yield a vector that can be compared with the PHOC vector derived from the textual query by cosine distance. In order to allow for a better comparison of Attribute-SVM score vectors and binary PHOC vectors, Almazán et al. use different methods for calibration. This includes Platt's Scaling [14] and different regression models. While Platt's Scaling is applied individually per SVM, regression also takes correlations between attributes into account.

Rusiñol et al. presented a method [7] that is very related to our approach. Building on top of their query-by-string approach [8], they introduced the query-by-speech paradigm for word spotting in historical documents. Feature representations from the visual and the audio domain are projected into a common subspace learned with LSI. In order to fuse these representations, they are concatenated to single vectors. Word images and audio signals are both represented by Spatial Pyramids using SIFT (Scale Invariant Feature Transform) descriptors for encoding local image regions and features based on cepstral coefficients for encoding short-time audio signals. They evaluate the retrieval performance of their method by using different text-to-speech synthesizers in known and unknown voice scenarios.

In this paper we present a method for segmentation-based word spotting with online-handwritten queries. Visual and online-handwriting features are embedded in the same subspace allowing for rapid distance-based retrieval. First of all, this new approach for a word spotting interface fits nicely with the recent trend towards more integrated and intuitive system interaction. Furthermore, there are two important methodological contributions. In order to represent online-handwritten words we introduce a new feature representation. Using standard geometrical features from online-handwriting recognition, Bag-of-Features Spatial Pyramids are applied in a conceptually equivalent manner as to features in the visual domain. As our system supports unknown writers, the expected variability in online-handwritten query representations is very high and therefore difficult to model in comparison to the static textual query representations given in query-by-string scenarios. In order to handle these variabilities, another important methodological contribution lies in the feature embedding pipeline. We propose to use distinct attribute embeddings for online-handwriting and visual features. This allows us to learn both representations individually from sample data with a framework of linear classifiers. As we will show in our experiments this is more suitable than using a single linear subspace model like LSI.

III. Method

The following section describes our approach for building a segmentation-based word spotting system for online-handwritten queries. Figure 1 shows an overview. The extension to segmentation-free scenarios is possible, as demonstrated for query-by-string word spotting in [15]. First, we will describe the visual descriptor used for word images and our novel descriptor used for representing online-handwritten words. Then, the attribute embedding is explained which uses PHOC vectors in order to calculate attributes for different feature domains. This method of building a common subspace for features of different domains builds on top of the work presented by Almàzan et al. [6]. Finally, we explain our approach of transforming feature representations of word images and online-handwritten trajectories into a common attribute space in detail.

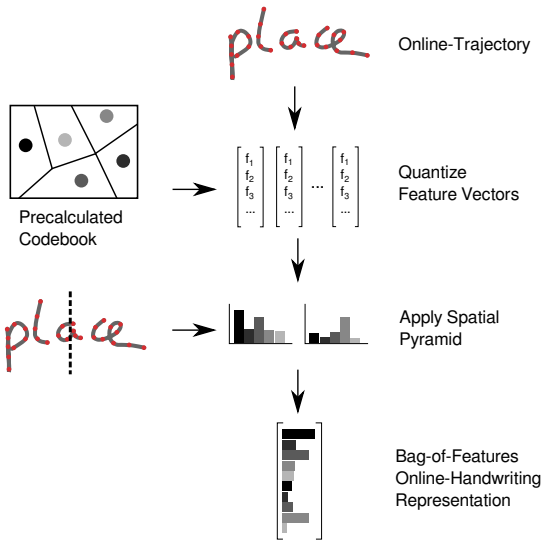

Fig. 2. Bag-of-Features online-handwriting representation. The feature extraction process is illustrated from top to bottom. An online-handwritten query is given by a sequence of trajectory points. For each point a vector of online-handwriting features is calculated and quantized with respect to a precalculated codebook of typical feature vector representatives. The online-trajectory is then divided into spatial regions composing a Spatial Pyramid. In this example, the regions are given by the left and right half of the trajectory. The quantized feature vectors are assigned to a region based on the coordinates of the represented point. For each region a histogram over the distribution of quantized vectors is built. Feature representatives are indicated in different tones of gray in the histograms. The final Bag-of-Features online-handwriting representation results from concatenating the histogram entries of all regions.

A. Visual Descriptor

In order to represent word images we use the Bag-of-Features approach (BoF, cf. [16]). A BoF image representation is an orderless collection (histogram) of local image features, in our case a dense grid of SIFT descriptors. The descriptors are quantized with respect to a visual vocabulary which is calculated from a large number of randomly chosen descriptors beforehand. We use a descriptor size of 40 pixels in a 5x5 grid. Each BoF image representation is extended by a Spatial Pyramid [17] to capture locality information. We separately normalize each region of the Spatial Pyramid with L2-norm. The final feature representation is built by concatenating histograms of all regions and normalizing the resulting vector by L2-norm.

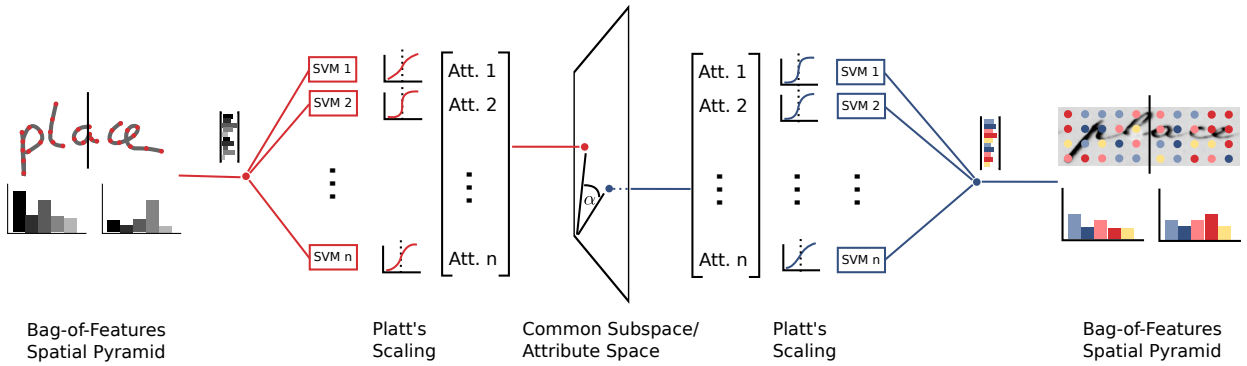

Fig. 3. Our approach for embedding the Bag-of-Features representations of a word image and an online-handwritten trajectory into a common attribute space. A separate set of SVMs is trained for both feature domains, where visual Attribute-SVMs are indicated in blue, and online-handwriting Attribute-SVMs are indicated in red. The scores are calibrated by means of Platt's Scaling, indicated by a sigmoid function for each SVM. The similarity between two attribute vectors is determined using the cosine distance, as shown by the angle in the common subspace.

B. Online-Handwritten Descriptor

We use a novel feature representation for online-handwritten trajectories based on the Bag-of-Features approach. The method is visualized in Figure 2. An online-handwritten trajectory, given by a sequence of trajectory points, is normalized with respect to slope, slant and height to reduce unwanted variability stemming from different writing styles. After resampling, a feature vector is calculated for each point of the trajectory. We use standard geometric features from online-handwriting recognition, i.e., writing direction, curvature, aspect ratio, curliness, lineness, slope and a context bitmap. For more details cf. [3], [18]. At this point each trajectory is represented by a sequence of feature vectors. In analogy to the BoF image representation, a large set of randomly chosen feature vectors is then clustered into a codebook (see diagram in Figure 2) of typical feature vector representatives.

The BoF online-handwriting representation for a new online-handwritten trajectory is built by calculating a feature vector for each point of the trajectory. These vectors are then quantized by their nearest neighbour in the codebook. In the Spatial Pyramid scheme, the quantized vectors are assigned to spatial regions, depending on the coordinates of their respective points. For each region a histogram is built, containing the occurrences for each codebook representative in that region. The final descriptor is formed by concatenating the L2-normalized histograms of each region and then normalizing the resulting vector with the L2-norm.

C. Attribute Embedding

We propose to use two separate attribute embeddings for visual and online-handwriting feature domains. In addition to the Attribute-SVMs for visual feature representations this leads to a second set of Attribute-SVMs to also determine the attributes for BoF online-handwriting representations described in Section III. The complete pipeline is shown in Figure 3. In the following we describe the handling of online-handwritten trajectories. The attribute embedding of word images works analogously, cf. [6].

In the training stage each trajectory is annotated with its transcription, for which the PHOC-vector can be calculated. As was the case for the visual feature representations of word images, we train an SVM for each PHOC attribute. A new online-handwritten trajectory is then scored by all these Attribute-SVMs to create the embedded attribute representation. Retrieval is done by measuring the distance between the embedded query and the embedded word images using cosine distance as the distance measure in the attribute space. While measuring vector distances we need to compare two attribute vectors composed of scores of independent SVMs. Since these scores have different dynamic ranges we evaluate the application of different means of calibration to the attribute vectors, cf. [6]. Firstly, Platt's Scaling [14] can be applied independently for each SVM of the two sets. Alternatively, linear regression or common subspace regression, cf. [6], is used in order to find a joint vector space for both embeddings while taking the correlations between different attributes into account.

IV. Evaluation

This section describes the experiments that were carried out in order to evaluate our proposed method of query-by-online-trajectory word spotting. First the datasets used for evaluation and our cross-validation protocol are presented. Then the results of the experiments are shown and compared to those of existing word spotting methods.

A. Datasets

Our method is evaluated on datasets of online-handwritten trajectories and historical document images.

The George Washington Dataset (GW, [4]) is a collection of 20 pages of handwritten letters by George Washington and his associates. It contains 4860 segmented word images which are annotated with their transcriptions. As a baseline for our query-by-online-trajectory word spotting approach we created a handwritten version of the George Washington Dataset, the George Washington Online dataset (GWO), which was written by a single writer on an Android Tablet. It contains

an online-handwritten trajectory for each of the 4860 word images mentioned above. The UNIPEN Dataset was published by the UNIPEN Foundation [19] and is composed of online-handwritten texts, segmented into lines, words and characters. We use a subset (denoted as *sta0*) from the whole UNIPEN corpus containing segmented online-handwritten word trajectories of 62 writers with an average of 400 word-trajectories per writer, amounting to a total of 27112 trajectories. This subset has a large amount of overlap in the vocabulary with the GW dataset.

B. Protocol

In order to validate and test our word spotting approach, we use a cross-validation protocol that is based on a widely used four-fold cross validation of the George Washington Dataset (e.g. [6], [8]). Our GWO dataset is divided into the same four folds, meaning each fold of the GW dataset contains exactly the same word occurrences and instances as the respective fold in the GWO dataset. In the experiments we optimize Spatial Pyramid configurations and BoF codebook sizes for the online-handwriting and document image domains. We use a fixed set of parameters for the SIFT descriptor and the grid size, similar to the configuration reported in [8], [10]. Starting from this basic setup, we conduct two different experiments.

Experiment 1: We test the query-by-online-trajectory word spotting performance in a scenario where the writer is already known to the system. In the training stage three folds of the GW dataset are used to learn the visual Attribute-SVMs and the corresponding three folds of the GWO dataset are used to train the online-handwriting Attribute-SVMs. The trained word spotting model is evaluated in the remaining GW and GWO fold, where each trajectory from the GWO fold is used as a query once and word images from the corresponding fold in the GW dataset are retrieved, cf. [7]. For the embedded attributes approach we evaluate the direct comparison of attribute scores (Att.) and three methods for calibrating these scores: Platt's scaling (Att.+Platt's), linear regression (Att.+Reg.) and common subspace regression (Att.+CSR). For comparison we also adopted the query-by-string/-speech approach by Aldavert, Rusiñol et al. [7], [8] for query-by-online-trajectory word spotting by replacing their textual/audio feature descriptor with our bag-of-online-features representation for calculating an LSI topic space.

Experiment 2: In order to test queries from an unknown writer, this experiment replaces the three folds of the GWO dataset in the training stage with trajectories from the UNIPEN dataset to train the online-handwriting Attribute-SVMs. The trajectories of the GWO dataset are only used for evaluation and are therefore unknown to the trained model. This experiment only uses word images and trajectories that represent words present in both UNIPEN and GW/GWO datasets. This reduces the amount of word images per page in the GW dataset to about 100 and the amount of trajectories for each writer in the UNIPEN dataset to about 110. We also report results for a variant in which all 27112 UNIPEN trajectories and the complete GW and GWO datasets are used. Note that in both variants we don't separate the UNIPEN dataset into folds since it is only used in training. We only evaluate the Att. and Att.+Platt's variant in this experiment since the other variants need corresponding word images and online-trajectories which represent the same word for calculating a subspace (LSA) or optimizing SVM scores (Att.+Reg., Att.+CSR). This requirement cannot be met when using the UNIPEN dataset.

TABLE I
QUERY-BY-ONLINE-TRAJECTORY RESULTS FOR EXPERIMENT 1.

Method	mAP in%
LSI	66.81
Att.	81.37
Att. + Platt's	79.99
Att. + Reg.	**86.49**
Att. + CSR	83.59

C. Results

The following results are reported in terms of mean Average Precision (mAP), a standard measure for evaluating retrieval performance. We are including in- and out-of-vocabulary queries. Please note that in our segmentation-based framework, retrieval lists always contain all word images that are relevant to the query.

The first experiment is designed to show the general feasibility of our query-by-online-handwriting word spotting approach. The results are presented in Table I. Att.+Reg. and Att.+CSR are the best options for our evaluation setup, which shows that SVM score calibration leads to better results than directly comparing scores or using the LSI method. Platt's scaling, that calibrates SVM scores independently, yields a mAP that is marginally lower than the Att. variant. The high mAP results for all attribute embedding variants are comparable to results of state-of-the-art query-by-string methods (see Table III). In contrast, the LSI method performs considerably worse than the attribute embedding.

The second experiment aims at evaluating a more realistic scenario of a word spotting model that is independent from the handwriting style used in queries, i.e. also unknown writers should be able to search for word images. Table II shows the parameter variation for the Att.+Platt's variant. The best result of 48.00% mAP was obtained with the optimal parameter choice of 2048 visual words (VW), 128 online-handwriting clusters (OHC) and a visual and online-handwriting Spatial Pyramid (VSP, OSP) of 2x1, 1x1 and 9x2, 3x2, respectively. Increasing the number of visual words to 4096 leads to the same result but also increases the overall size of the visual feature representation, depending on the size of the visual Spatial Pyramid, which produces a higher computational cost. For all other parameters, the configuration is at a local optimum. Using only direct attribute comparison (Att.) yields a result of 43.81% mAP with the same parameter configuration but using a 6x2, 2x1 online-handwriting Spatial Pyramid.

TABLE II
PARAMETER OPTIMIZATION FOR "ATT. + PLATT'S" IN EXPERIMENT 2.

VW	VSP	OHC	OSP	mAP in%
2048	2x1, 1x1	128	9x2, 3x2	**48.00**
1024	2x1, 1x1	128	9x2, 3x2	46.60
4096	2x1, 1x1	128	9x2, 3x2	48.00
2048	1x1	128	9x2, 3x2	45.60
2048	3x2, 2x1	128	9x2, 3x2	47.81
2048	2x1, 1x1	64	9x2, 3x2	47.60
2048	2x1, 1x1	256	9x2, 3x2	42.54
2048	2x1, 1x1	128	6x2, 2x1	43.59
2048	2x1, 1x1	128	9x2, 6x2	47.18

TABLE III
RESULTS FOR DIFFERENT CROSS-DOMAIN WORD SPOTTING QUERY TYPES.

Method	mAP in%
Query-by-string [6]	91.29
Query-by-speech [7] (known speaker)	51.25
Query-by-online-trajectory, proposed (known writer)	86.49
Query-by-speech [7] (unknown speaker)	15.98
Query-by-online-trajectory, proposed (unknown writer)	48.00

When running Experiment 2 with all words from the datasets, i.e., without removing words, that are not shared, the results are worse. This is to be expected, since the word spotting model has to cope with more variability both in terms of unknown words and unknown handwriting styles that have to be represented by the same models. Using the parameter configurations that were optimized using the smaller datasets, Att.+Platt's yields 21.71% mAP. Please note that a decrease in retrieval performance when using more training data is counter intuitive. However, our Attribute-SVMs are estimated with holistic feature representations for online-handwritten words and word images. This is different when estimating statistical sequence models on character level. Feature sequences are aligned with respect to the models and their training is more specific.

Finally, Table III shows a summary of recent cross-domain word spotting methods. Please note that retrieval results are not directly comparable, due to the very different properties of the query domains (string, speech and online-handwriting). In the known writer scenario we achieve remarkable results. In the unknown writer scenario our results are very promising given that the task is considerably harder than query-by-string word spotting.

V. CONCLUSION

In this paper we make three contributions to the area of word spotting systems. A new pen-based interface where online-handwritten words can be used as queries has been presented for the first time. Furthermore, there are two methodological contributions. Firstly, we propose to apply the well-known Bag-of-Features Spatial Pyramid approach to a sequence of feature vectors extracted from online-handwritten trajectories in order to form a single, highly discriminative descriptor. Secondly, this is, to the best of our knowledge, the first time that two separate SVM attribute embeddings have been applied for fusing features from different domains in a common subspace. In our experiments we achieve state-of-the art results in a writer-dependent scenario. In the case of a writer-independent scenario our results are compelling. With the growing presence of pen-based devices, especially smartboards, this is an important step towards a more interactive integration of querying word spotting systems.

REFERENCES

[1] J. Lladós, M. Rusiñol, A. Fornés, D. Fernández, and A. Dutta, "On the influence of word representations for handwritten word spotting in historical documents," *Int. Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 05, 2012.

[2] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A novel word spotting method based on recurrent neural networks," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 211–224, 2012.

[3] M. Liwicki and H. Bunke, "Hmm-based on-line recognition of handwritten whiteboard notes," in *Proc. of the Int. Workshop on Frontiers in Handwriting Recognition*. Suvisoft, 2006.

[4] "George Washington Papers at the Library of Congress, 1741–1799, Series 2, Letterbook 1, pages 270-279 and 300-309," Manuscript Division, Library of Congress, Washington, D.C., http://memory.loc.gov/ammem/gwhtml/gwhome.html.

[5] J. A. Rodríguez and F. Perronnin, "Label embedding for text recognition," in *Proc. of the British Machine Vision Conference*, 2013.

[6] J. Almazán, A. Gordo, A. Fornes, and E. Valveny, "Word Spotting and Recognition with Embedded Attributes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2552–2566, 2014.

[7] M. Rusiñol, D. Aldavert, R. Toledo, and J. Lladós, "Towards query-by-speech handwritten keyword spotting," in *Proc. of the Int. Conf. on Document Analysis and Recognition*, 2015.

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

[9] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *Int. Journal on Document Analysis and Recognition*, pp. 139–152, 2007.

[10] 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.

[11] 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.

[12] Y. Leydier, A. Ouji, F. LeBourgeois, and H. Emptoz, "Towards an omnilingual word retrieval system for ancient manuscripts," *Pattern Recognition*, vol. 42, no. 9, pp. 2089–2105, 2009.

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

[14] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[15] S. Ghosh and E. Valveny, "Query by string word spotting based on character bi-gram indexing," in *Proc. of the Int. Conf. on Document Analysis and Recognition*, 2015, pp. 881–885.

[16] S. O'Hara and B. A. Draper, "Introduction to the bag of features paradigm for image classification and retrieval," *Computing Research Repository*, vol. arXiv:1101.3354v1, 2011.

[17] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *Proc. of the IEEE Comp. Soc. Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 2169–2178.

[18] S. Jaeger, S. Manke, and A. Waibel, "Npen++: An On-Line Handwriting Recognition System," in *Proc. of the Int. Workshop on Frontiers in Handwriting Recognition*, 2000, pp. 249–260.

[19] "The UNIPEN on-line handwritten samples collection release #1," International Unipen Foundation, http://www.unipen.org/products.html.