High Performance Query-by-Example Keyword Spotting Using Query-by-String Techniques

Enrique Vidal
Universitat Politècnica de València
Valencia, Spain
Email: evidal@prhlt.upv.es

Alejandro H. Toselli
Universitat Politècnica de València
Valencia, Spain
Email: ahector@prhlt.upv.es

Joan Puigcerver
Universitat Politècnica de València
Valencia, Spain
Email: joapuipe@prhlt.upv.es

***Abstract*—Keyword Spotting (KWS) has been traditionally considered under two distinct frameworks: Query-by-Example (QbE) and Query-by-String (QbS). In both cases the user of the system wished to find occurrences of a particular keyword in a collection of document images. The difference is that, in QbE, the keyword is given as an exemplar image while, in QbS the keyword is given as a text string. In several works, the QbS scenario has been approached using QbE techniques; but the converse has not been studied in depth yet, despite of the fact that QbS systems typically achieve higher accuracy. In the present work, we present a very effective probabilistic approach to QbE KWS, based on highly accurate QbS KWS techniques which rely on models which need to be trained from annotated data. To assess the effectiveness of this approach, we tackle the segmentation-free QbE task of the ICFHR-2014 Competition on Handwritten KWS. Our approach achieves a *mean average precision* (mAP) as high as 0.715, which improves by more than 70% the best mAP achieved in this competition (0.419 under the same experimental conditions).**

I. Introduction

Perhaps billions of historical handwritten text documents remain untranscribed and current handwritten text recognition (HTR) technology is still far to offer sufficiently accurate automatic transcripts. Therefore, sheer amounts of important historical information remains practically inaccessible. Key Word Spotting (KWS) is one of the approaches which are being proposed to deal with this problem. KWS aims at determining locations on a text image or image collection which are likely to contain an instance of a queried word, without explicitly transcribing the image(s).

KWS is generally qualified as Query-by-Example (QbE) or Query-by-String (QbS), depending on whether the query word is specified by means of an example-image or as just a character string, respectively. Moreover, most of the techniques which have been proposed for KWS can be considered to belong to one of these two broad classes: *training-based* and *training-free*. Training-based KWS methods are generally based on statistical optical (and language) models and typically adopt the QbS paradigm [1], [2], [3], [4], [5], [6], [7], [8], [9]. On the other hand, most training-free techniques are based on direct template (image) matching and assume the QbE framework; see [10], [11], [12], [13], [14], [15], [16], among many other.

A possible disadvantage of QbS training-based methods is that they need a certain amount of transcribed page images to train on. But, after adequately trained, these methods can generally provide (much) greater *"precision-recall"* performance, measured for example in terms of *average precision* (AP). This AP superiority is explained not only by the well-know adaptation capability of (statistical) training, but also because a textual query can better represent the many forms a word can be rendered into an image, than a specific image corresponding only to one of these forms. In addition, most statistical models used for QbS KWS can easily incorporate word image contextual information through the use of language models.

Conversely, QbE training-free techniques generally provide lower AP, but they are often considered much more flexible; not only because no training is needed, but also because a query can be trivially specified even as a cropped (word) image region which the user does not know how to transcribe.

KWS approaches are additionally categorized into "segmentation-based" and "segmentation-free". The former assumes that the text images have been previously segmented into perfect word image regions. Because this is generally considered rather unrealistic in practice, only the segmentation-free assumption is considered in this work.

Some authors have proposed to further extend the flexibility of QbE techniques, letting these techniques to perform also QbS by means of rendering textual queries in the form of synthetic query images; see e.g., [12], [17], [18], [19], [20], [21], [22]. With a similar purpose, a different approach is followed in [23], where visual features are combined with textual information to support not only QbE, but also QbS (under the segmentation-based assumption).

In a symmetric way, given that QbS training-based methods can provide greater KWS performance, it is interesting to explore approaches to achieve QbE functionality using QbS methods. To our knowledge, this idea has never been studied in depth or developed so far.

We present a statistical approach to achieve the proposed functionality. Using this approach we consider the Bentham Dataset QbE KWS task proposed in the recent KWS ICFHR-2014 Competition[1] [24]. Under identical setting and constrains of the "segmentation-free" track, we improve the mean average precision (mAP) of the winner system by more than 70% relative (0.715 mAP, compared to 0.419).

The rest of the paper is organized as follows: First (Sec. II) we briefly introduce the training-based QbS methods used to develop our probabilistic QbE KWS approach. This approach is formally presented in Sec. III. Later we describe

[1] http://vc.ee.duth.gr/h-kws2014/

the experiments conducted to assess the performance of the proposed approach and the corresponding results (Sec. IV). Finally, the main conclusions and remarks of this work are drawn in Sec. V.

II. Training-based Query-by-String KWS

Here we outline the underlying training-based QbS methods, that will be used in the next section to develop the segmentation-based QbE approach. The details of these methods were originally presented in [4]. For each query word v and each text-line image region, x, represented by a sequence of feature vectors, a score $S(x, v)$ is defined. It measures the confidence that the keyword v is written in x or, in other words, how likely is that the "line image x is relevant for the keyword v". It is computed as:

$$S(x, v) \stackrel{\text{def}}{=} \max_{1 \leq i \leq n} P(v \mid x, i) \tag{1}$$

where $P(v \mid x, i)$ is the probability that the word v is present in the line image x at horizontal position i. As shown in [4], this posterior is approximated from the word graph (WG) of the line image.

A WG is a weighted directed acyclic graph encoding multiple transcripts, along with the corresponding probabilistic and segmentation information. WGs are obtained as a by-product of the standard Viterbi decoding, using a traditional HTR system [25]. Such a system is based on the so called "optical character models" (traditionally Hidden Markov Models) and a statistical Language model (generally an N-gram). These models are trained from adequate amounts of line text images accompanied with their corresponding transcripts.

This approach has been applied to achieve very good performance in several QbS KWS tasks [4], [9], [8].

Observe that the previous score can be interpreted as the probability of a Bernoulli distribution over a random variable $\mathcal{R}$, corresponding to the binary event "text line x is relevant for keyword v".

$$P(\mathcal{R} \mid x, v) \stackrel{\text{def}}{=} \begin{cases} S(x, v) & \mathcal{R} = 1 \\ 1 - S(x, v) & \mathcal{R} = 0 \end{cases} \tag{2}$$

If $P(\mathcal{R} \mid x, v)$ is well modeled, it will give scores close to 1 for relevant pairs of line images and (string) query keywords, and close to 0 for non-relevant events.

III. Extension to perform Query-by-Example KWS

A straightforward idea to implement a QbE system, based on the previous QbS approach, is to recognize each query image, by means of a standard handwritten word recognizer and use the obtained transcript as the query string for the QbS system. However, the probabilistic approach to KWS outlined in the previous section, leads to a formal, general solution from which this basic idea is just the simplest approximation.

As in the previous section, let v be a word identifier (string) and $P(\mathcal{R} \mid x, v)$ the probability that v, is written in an image (line) region x. Observe that v is unknown in the QbE scenario; instead, the query is specified by means of an example image, q. So our problem is to compute $P(\mathcal{R} \mid x, q)$.

The unknown transcript of q, i.e. v, can be introduced in $P(\mathcal{R} \mid x, q)$ as a hidden variable:

$$\begin{aligned} P(\mathcal{R} \mid x, q) &= \sum_v P(\mathcal{R}, v \mid x, q) \\ &= \sum_v P(\mathcal{R} \mid x, q, v) P(v \mid x, q) \end{aligned}$$

Now we reasonably assume that $P(v \mid x, q)$ is independent of x and that $P(\mathcal{R} \mid x, q, v)$ is independent of q given v, leading to:

$$P(\mathcal{R} \mid x, q) = \sum_v P(\mathcal{R} \mid x, v) P(v \mid q) \tag{3}$$

The first term of the sum in Eq. (3) is just the original QbS probability, for the string v in the image x, while the second one, $P(v \mid q)$, is the classification (posterior) probability of a word recognizer applied to the query image q. This becomes even more clear if the sum in Eq. (3) is approximated by its dominating term:

$$P(\mathcal{R} \mid x, q) \approx \max_v P(\mathcal{R} \mid x, v) P(v \mid q) \tag{4}$$

From this expression, the straightforward idea commented at the beginning of this section arises as a further approximation. To this end, the maximization problem (4) is naively solved by first maximizing the second term and then computing the first for a transcript $v^\star$ which maximizes the second, i.e.:

$$\begin{aligned} v^\star &= \arg\max_v P(v \mid q) \\ P(\mathcal{R} \mid x, q) &\approx P(v^\star \mid q) P(\mathcal{R} \mid x, v^\star) \end{aligned} \tag{5}$$

The computation cost involved in these equations can be reduced if the sum in v of Eq. (3), or the maximization in v of Eq. (4) and (5), is limited to the set of the n transcripts with highest $P(v \mid q)$. Moreover, it has been observed that this can also improve KWS performance (very) sightly. Therefore, a value of $n = 5$ was chosen to perform all the experiments reported in the next section.

IV. Experiments

A. Evaluation Measures

In order to evaluate the performance of the different KWS systems explored in this work, we used the evaluation software provided by the organizers of the ICFHR-2014 Competition on KWS [24]. More specifically, we mainly use the mean average precision (mAP) to compare the performance of the different systems.

Since we are in a segmentation-free scenario where the systems must provide a bounding box with the localization of each spotted keyword in each document page image, an additional measure is needed in order to decide whether the given bounding box is sufficiently correct. The organizers of the competition used the overlapping area between the reference bounding boxes and the detected ones, defined as IOA $= \frac{A \cap B}{A}$, where A and B are the bounding boxes given by the ground truth and the system, respectively. All detected keywords with an overlapping area greater than 0.7, are considered correct[2].

[2] We are aware that this overlapping measure can be easily fooled by letting the system to provide a bounding box equal to the whole page. However, for the sake of fair comparison, we did not take advantage of this shortcoming.

Finally, we compare our results to the official scoreboard of the ICFHR2014 Competition on KWS, using three other measures adopted during the competition to rank the contestants: *precision at 10* (P@10) computes the ratio of the number of detected keywords which are actually relevant, among the 10 best detected keywords of each query image. They also reported results using the *normalized discounted cumulative gain* (NDCG) metric. The NDCG metric is used when non-binary relevance judgment is used to evaluate the performance of the KWS system (i.e. the relevance of a keyword is a real number in the range $[0 - 1]$). Two NDCG scores are given: one assuming a binary relevance judgment and the second assuming non-binary judgment. Further details about these evaluation measures are explained in [24].

B. Dataset

The test dataset, is exactly the same used in the ICFHR-2014 Competition on KWS. It consists of 50 page images of handwritten manuscripts from the Bentham collection, written by Jeremy Bentham (1748-1832) and his secretarial staff.

To train and validate the required HMM and bi-gram models for the proposed KWS approach both transcribed text images and additional plain text data where used. The training transcribed images (also a subset of the Bentham collection) where taken from the original training and validation sets of the HTRtS ICFHR-2014 Competition [26]. Upon inspection, these two partitions happened to include some of the pages in the above mentioned test set of the KWS ICFHR-2014 Competition. Therefore, to prevent unfair "training-on-testing", we excluded them from training and validation sets. The resulting training and validation partitions consist of 300 and 50 page images, respectively, along with their transcripts and manually segmented line image regions. Statistics of these partitions are shown in Table I.

The transcripts of the training text images, along with additional plain text data, were used to train an interpolated bi-gram model [27]. The additional text was collected from several sources, including previously transcribed Bentham texts and selected parts of general texts from ECCO-TCP Eighteenth Century Texts [28], [27]. Of course, it was carefully verified that *no* chunk of text appearing in the text images was included in this training and validation text corpora. The "Text Data" rows of Table I provide relevant figures of this textual data set.

TABLE I: Image and text training and validation partitions.

		Training	Validation
Image Data	Pages	300	50
(Bentham page images)	Lines	8 019	1 291
	Running chars.	373 604	61 859
	Character set	93	84
Text Data	Running words	10 855 571	12 221
(Bentham + other texts)	Lexicon size (words)	78 311	2 602

C. Set of Keywords to be Spotted

The query set is the same used in the ICFHR-2014 Competition on KWS, consisting of 290 word image queries, obtained from the Bentham corpus. All query images correspond to keywords that were written more than 5 times in the considered test partition, and had a length greater than 6 characters. The query images come from several writers, have different writing styles, font sizes, and noise. Query images have to be spotted in images of pages written by the same or different writers and/or exhibiting different writing variations. Fig. 1 shows three query images corresponding to the same keyword to illustrate the query set variability.

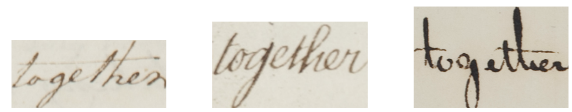

Fig. 1: Examples of three query images showing the variations of the same keyword, "together".

D. Experimental Set-up

We used the line segmentation given in the training set and extracted a 24 feature vector sequence from each text line, following [29]. Using the Baum-Welch algorithm, a left-to-right HMM was trained for each of the most frequent 87 characters appearing in the training set, such as lowercase and uppercase letters, symbols, special abbreviations, etc. The HMM metaparameters used in the final HMM training were optimized on the 50 pages of validation partition. As a result, all character HMMs were set up to have 6 states with 128 Gaussian mixtures per state. Additionally a bi-gram language model was trained, with Kneser-Ney smoothing, on the external Bentham text corpus (see Table I). Language model metaparameters such as *grammar scale factor* and *word insertion penalty* were also tuned on the validation set.

Then, we extracted the text line images from the test page images following two approaches: manual line segmentation and an automatic line segmentation based on full-page horizontal projection profile (HPP) [30]. For each line image, a WG with a maximum input degree of 15 is obtained and then used to compute the QbS scores (see Sec. II).

We test our KWS approaches corresponding to Eq. (3-5) in three different scenarios, which are increasingly close to the QbE KWS setting:

1) Both the line segmentation of page images and the transcripts of the query images are given (this would correspond to conventional line-level QbS KWS but, in addition, the horizontal positions of the spotted words within the spotted lines must be determined).
2) Line segmentation is given, but only raw, untranscribed query images are used. (this would correspond to QbE KWS with pre-segmented lines).
3) Line segmentation is obtained automatically and only untranscribed query images are used (this is the most standard segmentation-free QbE KWS setting, as used in the ICFHR-2014 KWS Competition).

In addition, under Scenario 3, we also try a totally naive baseline approach in which the best transcripts obtained by the query image word recognizer are used to plain-text search on the (error-prone) raw text (and the corresponding line image word segmentations) provided by a state-of-the-art handwritten text recognizer for automatically segmented lines. This recognizer was based on the same language model and character HMMs used here to compute $P(v \mid x, i)$ in Eq. (1). This will be referred to as *"raw transcripts"* below.

A minor, but crucial detail, is that all the approaches described in Sec. II and III provide relevance scores for whole line images. However, the ICFHR-2014 KWS competition rules require also to determine the bounding boxes of the spotted words. To this end, we simply obtain the highest likelihood segmentation of the word v in the line image x. This effectively means that if a keyword appears twice in a line, only one match will be retrieved. In practice, this did not affect the results significantly, since most of the test lines only contain one instance of the queried keyword[3].

E. Results

The approximations to $P(\mathcal{R} \mid x, q)$ given by Eq. (3-5) were used to evaluate the mAP obtained in the different scenarios described in Sec. IV-D. As expected, Eq. (3) achieved the best results; but only insignificantly better than those of Eq. (4) which, in turn, were only slightly better than those of Eq (5). For the sake of conciseness, only the results of Eq. (4) are reported in Table II, along with the *raw transcripts* naive baseline for *Scenario 3* discussed in Sec. IV-D.

TABLE II: Mean average precision in different scenarios, with $P(\mathcal{R} \mid x, q)$ approximated by Eq. (4).

Scenario 1	Scenario 2	Scenario 3	3 + Raw transcripts
0.863	0.865	0.715	0.547

It may seem a contradiction that the mAP achieved in Scenario 1 is (slightly) lower than in Scenario 2. Nevertheless it is perfectly understandable since, in Scenario 2, $P(\mathcal{R} \mid x, q)$ is affected by both $P(v \mid q)$ and $P(\mathcal{R} \mid x, v)$. Thus, for instance, if the given transcript v is not clearly written in an image line x, then in Scenario 1 $P(\mathcal{R} \mid x, v)$ can become close to 0. However, one (or more) similar word(s) v' may exist such that both $P(\mathcal{R} \mid x, v')$ and $P(v' \mid q)$ are significant; then Scenario 2, using Eq. (4) will give a larger $P(\mathcal{R} \mid x, q)$, thereby achieving better precision and recall. Of course, this will not happen very often, and will have, in general, tiny effects over the mAP, as it happens in this experiment, probably because of the small query set used in this task (290 query images, over 50 pages). In general, it is expected than the closer we are to the pure QbE case, the lower mAP is achieved, since the KWS system has to deal with more uncertainty. The difference between Scenario 2 and Scenario 3 support this hypothesis. Finally, as expected, it can be seen that the naive *raw transcripts* method actually falls short of providing competitive performance.

Table III shows the official scoreboard of the ICFHR-2014 Competition on KWS, for the Bentham dataset, and the results achieved by the method presented in this work (i.e. Scenario 3, using Eq. (3)). At the expense of requiring training data, our proposed QbE KWS approach is highly superior to any of the others, regardless the evaluation metrics used.

Fig. 2 shows mean Recall-Precision curves of our methods and that of the ICFHR-2014 KWS Competition winner system.

[3]Rather than the best match, the N-best matching positions of a word in a line can be trivially obtained. But, in practice, this is completely irrelevant since real use only demands finding the lines where words are likely to appear (this also makes superfluous to find accurate bounding boxes).

TABLE III: Comparison between the official scoreboard of the ICFHR-2014 Competition on KWS and this work.

	P@5	NDCG (bin.)	NDCG	mAP
Team 1	0.611	0.640	0.657	0.419
Team 3	0.568	0.518	0.536	0.372
Team 4	0.341	0.363	0.376	0.209
Team 5	0.550	0.513	0.531	0.347
This work	**0.879**	**0.822**	**0.823**	**0.715**

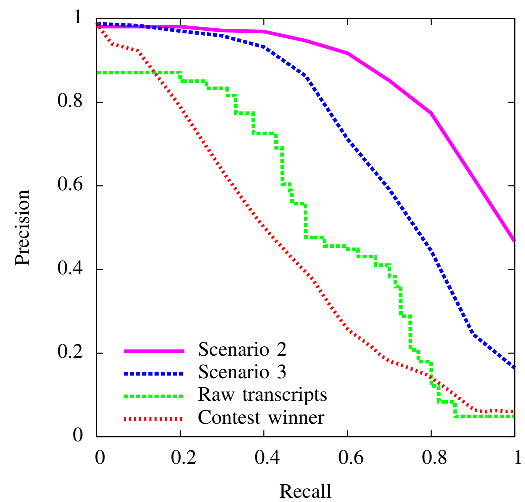

Fig. 2: Recall-Precision curves achieved in this work, along with that of the winner of the ICFHR-2014 KWS Competition.

V. Remarks and Conclusions

We presented an approach to tackle the query-by-example (QbE) scenario using a probabilistic KWS system designed under the query-by-string (QbS) framework.

Not only we have shown that QbE can be properly approached from the QbS perspective, but also that much higher performance can be obtained in this way, assuming that an adequate amount of training data is available. We tested our system for increasing levels of "difficulty", going from a QbS scenario to full QbE, under different levels of supervision.

We compared our system under the same conditions than other four systems designed for QbE, using the evaluation data and tools of the ICFHR-2014 Competition on Handwritten KWS. The approach presented in this paper showed to be much better than any of the contestants, for a variety of evaluation metrics. More specifically, it provided more than a 70% relative increase in the mean average precision (mAP) metric, with respect to the winner team of the competition (0.715 vs. 0.419).

Moreover, much larger improvements are possible by avoiding or reducing line segmentation errors. As shown by results of scenario 2 (which is line-segmentation error-free), the mAP of the proposed QbE approach using Eq. (4) jumps up to 0.864, or a relative mAP increase of 106% over the best result so far. In our upcoming work, we will try to completely avoid these segmentation errors by extending our KWS approaches so as to avoid explicit line segmentation.

It is important to remind that the dramatic improvements achieved by the proposed methods came at the cost of requiring sufficiently large image and text training data sets. In the present experiments we have used *all* the relevant training data which was easily available. However, perhaps the most important question that arises from this work is how much training data is really enough to effectively train a KWS system based on the proposed approaches, and whether these data must come from the same collection of documents, or maybe the required models can be transferred from a different task. This, and other questions, need further research to be answered, but this work is a successful step towards.

ACKNOWLEDGMENTS

The authors are very thankful to Jesse De Does and the INL team, for sharing the interpolated bigram language model used in the experiments and to the organizers of the ICFHR-2014 KWS Competition, for providing the data, evaluation tools and detailed information about the competition.

Work partially supported by the Spanish MEC under FPU grant FPU13/06281 and under the STraDA research project (TIN2012-37475-C02-01), by the Generalitat Valenciana under the grant Prometeo/2009/014, and through the EU 7th Framework Programme grant tranScriptorium (Ref:600707).

REFERENCES

[1] S. Wshah, G. Kumar, and V. Govindaraju, "Script independent word spotting in offline handwritten documents based on hidden markov models," in *Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on*, 2012, pp. 14–19.

[2] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke, "A Novel Word Spotting Method Based on Recurrent Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 211 –224, Feb. 2012.

[3] A. Fischer, V. Frinken, H. Bunke, and C. Suen, "Improving hmm-based keyword spotting with character language models," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, Aug 2013, pp. 506–510.

[4] A. H. Toselli, E. Vidal, V. Romero, and V. Frinken, "Word-graph based keyword spotting and indexing of handwritten document images," Universidad Politcnica de Valencia, Tech. Rep., 2013.

[5] A. Toselli and E. Vidal, "Fast HMM-Filler approach for key word spotting in handwritten documents," in *12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 501–505.

[6] G. Kumar and V. Govindaraju, "Bayesian active learning for keyword spotting in handwritten documents," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 2041–2046.

[7] J. Almazan, A. Gordo, A. Fornes, and E. Valveny, "Handwritten word spotting with corrected attributes," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, Dec 2013, pp. 1017–1024.

[8] J. Puigcerver, A. H. Toselli, and E. Vidal, "Word-graph-based handwriting keyword spotting of out-of-vocabulary queries," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 2035–2040.

[9] A. H. Toselli and E. Vidal, "Word-graph based handwriting key-word spotting: Impact of word-graph size on performance," in *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*. IEEE, 2014, pp. 176–180.

[10] A. Kolcz, J. Alspector, M. Augusteijn, R. Carlson, and G. Viorel Popescu, "A Line-Oriented Approach to Word Spotting in Handwritten Documents," *Pattern Analysis & Applications*, vol. 3, pp. 153–168, 2000, 10.1007/s100440070020.

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

[12] T. Konidaris, B. Gatos, K. Ntzios, I. Pratikakis, S. Theodoridis, and S. J. Perantonis, "Keyword-guided word spotting in historical printed documents using synthetic data and user feedback," *Int. Journal of Document Analysis and Recognition*, vol. 9, no. 2-4, pp. 167–177, 2007.

[13] B. Gatos and I. Pratikakis, "Segmentation-free word spotting in historical printed documents," in *10th International Conference on Document Analysis and Recognition, 2009. ICDAR'09*. IEEE, 2009, pp. 271–275.

[14] M. Rusinol, D. Aldavert, R. Toledo, and J. Llados, "Browsing Heterogeneous Document Collections by a Segmentation-Free Word Spotting Method," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2011, pp. 63 –67.

[15] A. Fornés, V. Frinken, A. Fischer, J. Almazán, G. Jackson, and H. Bunke, "A keyword spotting approach using blurred shape model-based descriptors," in *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*. ACM, 2011, pp. 83–90.

[16] D. Fernández, J. Lladós, and A. Fornés, "Handwritten word spotting in old manuscript images using a pseudo-structural descriptor organized in a hash structure," in *Pattern Recognition and Image Analysis*. Springer, 2011, pp. 628–635.

[17] A. Kumar, C. Jawahar, and R. Manmatha, "Efficient search in document image collections," in *Computer Vision–ACCV 2007*. Springer, 2007, pp. 586–595.

[18] A. Kesidis, E. Galiotou, B. Gatos, A. Lampropoulos, I. Pratikakis, I. Manolessou, and A. Ralli, "Accessing the content of greek historical documents," in *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*. ACM, 2009, pp. 55–62.

[19] J. A. Rodriguez-Serrano and F. Perronnin, "Handwritten word image retrieval with synthesized typed queries," in *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*. IEEE, 2009, pp. 351–355.

[20] R. Jain and C. Jawahar, "Towards more effective distance functions for word image matching," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*. ACM, 2010, pp. 363–370.

[21] A. L. Kesidis, E. Galiotou, B. Gatos, and I. Pratikakis, "A word spotting framework for historical machine-printed documents," *International Journal on Document Analysis and Recognition*, vol. 14, no. 2, pp. 131–144, 2011.

[22] J. A. Rodriguez-Serrano and F. Perronnin, "Synthesizing queries for handwritten word image retrieval," *Pattern Recognition*, vol. 45, no. 9, pp. 3270–3276, 2012.

[23] D. Aldavert, M. Rusinol, R. Toledo, and J. Lladós, "Integrating visual and textual cues for query-by-string word spotting," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 511–515.

[24] I. Pratikakis, K. Zagoris, B. Gatos, G. Louloudis, and N. Stamatopoulos, "ICFHR 2014 Competition on Handwritten KeyWord Spotting (H-KWS 2014)," 2014.

[25] S. Ortmanns, H. Ney, and X. Aubert, "A word graph algorithm for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 11, no. 1, pp. 43–72, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/B6WCW-45N4RJB-N/2/0f22c5b07ee9378da100a928907910e7

[26] J. A. Sánchez, V. Romero, A. H. Toselli, and E. Vidal, "ICFHR2014 Competition on Handwritten Text Recognition on tranScriptorium Datasets (HTRtS)," 2014.

[27] J. Tanha, J. D. Does, and K. Depuydt, "An intelligent sample selection approach to language model adaptation for hand-written text recognition," in *14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2014.

[28] M. Moyle, J. Tonra, and V. Wallace, "Manuscript transcription by crowdsourcing: Transcribe bentham," *LIBER Quarterly*, vol. 20, no. 2, 2011.

[29] M. Kozielski, P. Doetsch, and H. Ney, "Improvements in rwth's system for off-line handwriting recognition," in *International Conference on Document Analysis and Recognition*, Washington, DC, USA, Aug. 2013, pp. 935–939.

[30] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *International Journal on Document Analysis and Recognition*, vol. 9, pp. 123–138, April 2007.