Pattern Analysis and Applications (2019) 22:23–32
https://doi.org/10.1007/s10044-018-0742-z

ORIGINAL ARTICLE

Probabilistic multi-word spotting in handwritten text images

Alejandro H. Toselli[1] · **Enrique Vidal**[1] · **Joan Puigcerver**[1] · **Ernesto Noya-García**[1]

Received: 29 September 2017 / Accepted: 20 July 2018 / Published online: 3 August 2018

Abstract
Keyword spotting techniques are becoming cost-effective solutions for information retrieval in handwritten documents. We explore the extension of the single-word, line-level probabilistic indexing approach described in our previous works to allow for page-level search of queries consisting in Boolean combinations of several single-keywords. We propose heuristic rules to combine the single-word relevance probabilities into probabilistically consistent confidence scores of the multi-word boolean combinations. An empirical study, also presented in this paper, evaluates the search performance of word-pair queries involving *AND* and *OR* Boolean operations. Results of this study support the proposed approach and clearly show its effectiveness. Finally, a web-based demonstration system based on the proposed methods is presented.

Keywords Handwritten text processing · Keyword spotting · Multi-word Boolean queries · Image processing · Pattern recognition

1 Introduction

In recent years, large collections of historical handwritten documents are being scanned into digital images, in order to make them available through web sites of libraries and archives all over the world. However, the wealth of information conveyed by the text captured in these images remains largely inaccessible. Transcribing such documents by paleography experts is usually very expensive. Consequently, to exploit and make profit of such mass-digitization efforts, affordable information retrieval methods are required which allow the users to accurately and efficiently search for textual contents in large collections of *untranscribed* handwritten text images. This is one of the goals of projects such as HIMANIS[1] [3] and READ,[2] where probabilistic indexing methods based on *line-oriented word-segmentation-free Keyword Spotting* (KWS) are being developed [27, 29]. These methods rely on the same models used in handwritten text recognition (HTR), such as *recurrent neural networks* (RNNs) [7, 11, 24] or *hidden Markov models* (HMM) [2, 26, 30] for optical modeling, and *N*-grams for language modeling. Using these models, probabilistic word indices are built, assuming the finest search unit is the line image; that is, whole line images are analyzed to determine the degree of confidence that each given keyword appears in the image.

However, for searching in large collections involving millions of page images, line-level indexing can be less than adequate. The storage space required for the fine-grained line-level indices might become prohibitive and, on the other hand, a coarser, page-level search can be more than enough in most applications. Moreover, aiming at practical applications involving the search of general information in large image collections, we consider queries consisting in *Boolean combinations* of *multiple words*.

Boolean multi-word search can be implemented using any of the single-word KWS systems cited above. First, each word of the query is spotted separately, obtaining a set of spots (that is, lines or regions) in which each word is likely to appear above the specified confidence threshold. Then, set union, intersection and complement operations are applied to the resulting single-word spot sets to obtain the resulting

✉ Alejandro H. Toselli
ahector@prhlt.upv.es

Enrique Vidal
evidal@prhlt.upv.es

Joan Puigcerver
joapuipe@prhlt.upv.es

Ernesto Noya-García
noya.ernesto@gmail.es

[1] PRHLT Research Centre, Universitat Politècnica de València, Camino de Vera S/N, 46022 Valencia, Spain

[1] https://www.himanis.org
[2] https://read.transkribus.eu

set of spots of the given Boolean combined query. Yet, this still needs proper ways to combine the single-word confidence scores into the overall score of the Boolean query and check whether the combined score is higher than the given threshold.

An example of this approach is [21], which presents a (segmentation-based) KWS approach for multi-word queries formulated only with *AND/OR* Boolean operations. However, this approach has two main drawbacks: First, it requires a (perfect in the experiments of [21]) segmentation of all the images into individual words, which is obviously not affordable in practice for large image collections. And second, the implementation of the *AND* operation is inconsistent in probabilistic terms.

Clearly, only if the spotting scores are well normalized and probabilistically sound, we can follow standard probability laws to study how to consistently and adequately combine these scores. This is the idea we follow in all our works on KWS. Here, we extend the line-level indexing approach described in [27, 29] to build probabilistic word indices at the page level. In addition, we explore the feasibility of Boolean combination of single-word queries by introducing heuristic, albeit probabilistically consistent confidence score combination rules. Empirical results for page-level *AND* and *OR* word-pair Boolean queries are reported which support the consistency and usefulness of the proposed approach. This paper complements the work presented in [18] by reporting a new, larger empirical study aimed to measure the precision-recall performance of the different types of multi-word queries in a comparable way. It also includes a description of a real handwritten information retrieval system implemented using the proposed methods.

The rest of the paper is organized as follows. Section 2 overviews the probabilistic framework of single-keyword indexing and Sect. 3 introduces the probabilistic spotting scores proposed to support multi-word queries with Boolean operators. Dataset, evaluation measures, query selection and experimental set-up are presented in Sect. 4, and the empirical results are reported in Sect. 5. Section 6 outlines a demonstration system built following the proposed approach. Finally, Sect. 7 summarizes the work presented, draws conclusions and outlines future work.

2 Single-word probabilistic indexing

The indexing approach proposed in this work follows the KWS ideas originally presented in [27]. Here, a probabilistic word index is built at the page level. Let a page image, $\mathbf{x}$, be represented by their L text line images, $\mathbf{x}_1, \ldots, \mathbf{x}_L$. In turn, let each text line $\mathbf{x}_l$ be described as a "frame sequence": $\mathbf{x}_l = x_{l1}, x_{l2}, \ldots, x_{lJ_l}$, A frame is a subimage of $\mathbf{x}_l$ composed of some (or maybe one) contiguous line image columns, or a feature vector extracted from such subimage (typically used with HMMs [23]). For each query word v and each page image $\mathbf{x}$, a score $S(\mathbf{x}, v)$ is obtained which measures how likely is the event "keyword v is written in $\mathbf{x}$", or re-phrased as "page image $\mathbf{x}$ is relevant for keyword v". This score is computed as:[3]

$$S(\mathbf{x}, v) \stackrel{\text{def}}{=} \max_{1 \le l \le L} \max_{l \le j \le J_l} P(v \mid \mathbf{x}, l, j) \tag{1}$$

where $P(v \mid \mathbf{x}, l, j)$, called *line-level frame word posterior*, is the probability that the word v is present in the page image $\mathbf{x}$ at line l and frame position j.

As shown in [27], the line-level frame word posteriors required for Eq. (1) can be accurately and efficiently computed for each word v in a given lexicon or vocabulary V, using the same kind of optical, lexical and language statistical models as those used in HTR. In most previous works, N-grams and HMMs/RNNs have been used for language and character optical modeling, respectively. These models are trained from moderate amounts of training images accompanied by the corresponding transcripts using well known statistical estimation techniques [11, 12]. The lexicon, V, on the other hand, is also obtained from the training transcripts and possibly expanded with additional words obtained from other relevant texts, if they are available. Using these models, $P(v \mid \mathbf{x}, l, j)$ is computed for each l using a word-lattice, which is in turn obtained through an extension of the conventional process used to decode $\mathbf{x}_l$ into its best transcript [27].

Since $P(v \mid \mathbf{x}, l, j)$ is a well-defined discrete probability function, the score $S(\mathbf{x}, v)$ given in Eq. (1) can be properly used to define the following Bernoulli distribution:

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

where the random variable R represents the event "page image $\mathbf{x}$ is relevant for keyword v". In order to explicitly assume this probabilistic meaning of $S(\mathbf{x}, v)$, from now on we will refer to it as $P(R = 1 \mid \mathbf{x}, v)$, or simply $P(R \mid \mathbf{x}, v)$.

To produce the probabilistically index of a page image $\mathbf{x}$, $P(R \mid \mathbf{x}, v)$ is computed for all $v \in V$ and non-negligible values are retained. This (moderately intensive [27], but *off-line*) computation is carried out for all the images of the collection to be indexed. The resulting values are stored into an adequate database or data structure, $\mathscr{D}$, along with geometrical information about the location and size of v within $\mathbf{x}$. Then for a given (single-keyword) query w, $\mathscr{D}$ is searched

[3] In practice, the values of l and j associated to the maximum are also obtined. To deal with multiple instances of the same word in $\mathbf{x}$, not only a single maximum but the N highest maxima are actually retained.

for those entries $\mathbf{x}$ such that $P(R \mid \mathbf{x}, w) > \tau$, where τ is a threshold more or less explicitly specified by the user along with w itself. The off-line indexing phase avoids heavy computations during user's query look-up and permits extremely fast query processing.

3 Multi-keyword spotting

To simplify notation, in this section $P(R \mid \mathbf{x}, v)$ will be just denoted as $P(R_v \mid \mathbf{x})$. Moreover, we restrict the discussion to a fixed page image $\mathbf{x}$, so it can be dropped from the formulation. This way, $P(R \mid \mathbf{x}, v)$ becomes just $P(R_v)$.

We are interested in queries that are Boolean combinations of several keywords, $v_1, ..., v_M$, using the three basic Boolean operators: *OR*, *AND* and *NOT*, respectively denoted as "$\vee$", "$\wedge$" and "$\neg$". The relevance of $\mathbf{x}$ for an m-fold *AND* query is then written as $R_{v_1} \wedge R_{v_2} \cdots \wedge R_{v_M}$, or just $R_1 \wedge R_2 \cdots \wedge R_M$, for the sake of further simplifying notation. Similarly, the event for an *OR* query is denoted as $R_1 \vee R_2 \cdots \vee R_M$.

Computing the probability of events associated with arbitrarily complex combinations of these Boolean operators can become very complex and, moreover, even for the simplest cases, the probabilities of conditional dependencies (which can hardly be ignored) are needed. Therefore, in this paper, we propose convenient, efficiently computable approximations, based on the early work of Boole and Fréchet [5, 9, 10], and we assess their suitability for multi-word KWS through empirical tests presented in Sects. 4 and 5. These approximations are:

$$P(R_1 \wedge R_2 \cdots \wedge R_M) \approx \min(P(R_1), P(R_2), \ldots, P(R_M)) \quad (3)$$

$$P(R_1 \vee R_2 \cdots \vee R_M) \approx \max(P(R_1), P(R_2), \ldots, P(R_M)) \quad (4)$$

In addition, the relevance probability of the *NOT* operator applied to a Boolean query combination, B, is computed as:

$$P(\neg B) = 1 - P(B) \quad (5)$$

Using these equations, the (approximate) relevance probability of any arbitrary Boolean combination of single-keyword queries can be easily and very efficiently computed. For example, to search for image regions containing both the words *"cat"* and *"dog"* but none of the words *"mouse"* or *"rabbit"* the relevance probability will be computed as:

$$P(R_1 \wedge R_2 \wedge \neg(R_3 \vee R_4)) \approx \min(P(R_1), P(R_2), (1 - \max(P(R_3), P(R_4))))$$

where the events R_1, R_2, R_3 and R_4 correspond to the keywords *"cat"*, *"dog"*, *"mouse"* and *"rabbit"*, respectively.

4 Experiments

To assess the effectiveness of the proposed multi-word query spotting approach, several experiments were carried out. The dataset, the evaluation measures and the experimental set-up are presented in this section.

4.1 Dataset

The whole set contains more than 80,000 images of manuscripts written by the renowned English philosopher and reformer Jeremy Bentham (1748–1832) and his secretarial staff [6]. Page images of the Bentham collection (see examples in Fig. 1) generally require non-trivial preprocessing and layout analysis to deal with noisy and/or faint writing, marginal notes, stamps, skewed images, lines with different slope in the same page, variable slant, inter-line text, etc.

From the Bentham data currently available, a dataset of 433 page images is used in this work. This dataset contains nearly 100, 000 running words from a vocabulary of more than 9000 different words. Table 1 summarizes the basic statistics of this dataset.

As discussedt in Sect. 2, our indexing models need to be trained from data. Therefore, the dataset was divided into three subsets for training, validation and test, respectively encompassing 350, 50 and 33 images. Since it was not possible to accurately identify the writers in all cases, the pages were shuffled before distributing them over these three subsets. This means that some writers can appear both in the training and in the test sets. For each page image, text line regions were automatically obtained and manually revised. Note that a non-negligible number of these regions are short lines; for example, 9.5% of them contain just one word.

This dataset is exactly the same used in the ICFHR'14 handwritten text recognition competition [1], which is also part of the dataset used in the ICFHR'14 KWS competition [19]. On the other hand, it was employed as the training set of the ICDAR'15 KWS competition [20], with a test set of 70 pages. This test set was larger than the one used in this work, but the query set (243 single words) was much smaller (see Sect. 4.2). Finally, the size of the dataset used here is comparable to that of other standard datasets used for KWS benchmarking: George Washington [15] (20 pages), IAMDB [17] (1539 small form images), Parzival [8] (45 pages), etc.

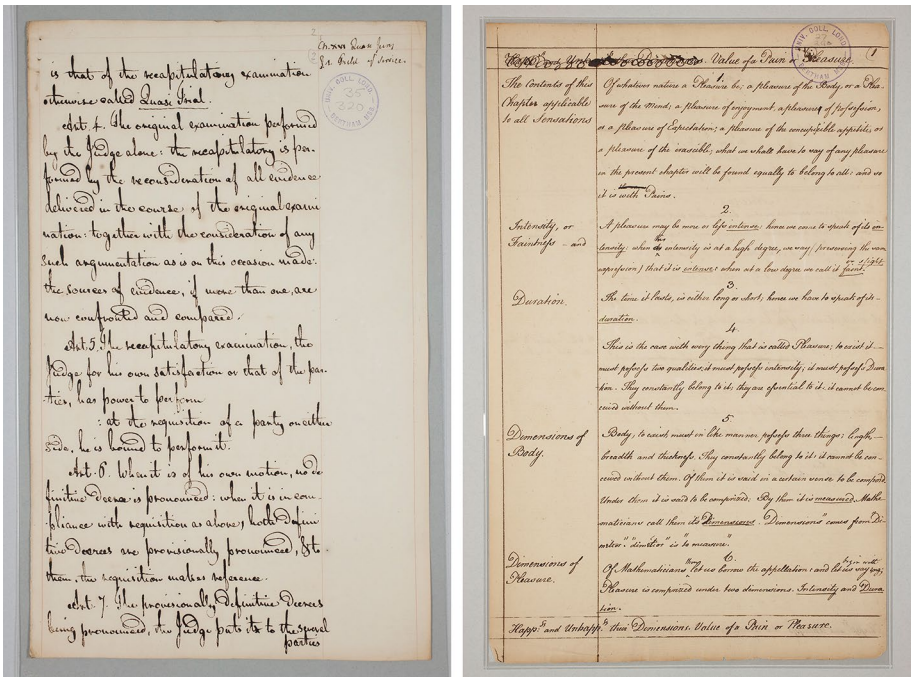

Fig. 1 Examples of Bentham page images

Table 1 Basic statistics of the Bentham dataset used in this work

Number of	Training	Validation	Test	Total
Pages	350	50	33	433
Lines	9198	1415	860	11,473
Running tokens	76,675	11,588	6955	95,218
Different tokens	12,220	3265	2282	13,978
Character set size	86	86	86	86
Running words	86,075	12,962	7868	106,905
Vocabulary size	8658	2709	1946	9716
OOV running words (%)	–	6.62	5.30	–
OOV words (%)	–	25.14	19.37	–

A *token* is any non-blank sequence of characters, while a *word* is assumed not to contain punctuation marks and each punctuation mark is considered a "word" by itself. "OOV" means "out of vocabulary"

Table 2 Basic statistics of the *SINGLE*, *AND* and *OR* pools of queries generated

	Query type	Total	Pertinent	$r_{\max}$
Queries	*Single*	3293	674	3.89
	OR	5, 420, 278	1992, 007	1.72
	AND	5, 420, 278	11, 784	458.97
Query events	*Single*	108, 669	836	128.99
	OR	178, 869, 174	2, 739, 674	64.29
	AND	178, 869, 174	12, 438	14, 379.86

The maximum ratios, $r_{\max}$, between non-pertinent and pertinent queries and events for each type of query are also reported

4.2 Query selection

As commented in Sect. 1, the empirical study presented in this paper explores the performance of handwritten text KWS for queries composed of one or two keywords, combined using the two Boolean operators *AND* and *OR*. Both for single and word-pair queries, the individual words were selected from a subset of *training* words.[4]

This subset, referred to as S, was composed of 3293 words whose frequency of occurrence in the training partition ranges from 2 to 10. This avoids including most *stop words* (generally with word frequencies greater than 10) and also many (singleton) words that are unlikely to appear in the test partition. In order to test word-pair *AND/OR* queries, a set, S2, of all the 5, 420, 278 pairs of different words in S was also generated.

Despite the selection criteria adopted, only a relatively small subset of 674 words from S does appear in the test images. The corresponding single-word queries are called *"pertinent"*, while all the other queries are called *"non-pertinent"*. Similarly, not all the word-pairs in S2 are pertinent for *AND* and *OR* query types. The total (maximum) number of pertinent queries which can be composed for each query type are reported in Table 2, along with the other figures mentioned above. The table also shows the corresponding

[4] In many works on KWS, query sets are selected from the *test* data instead.

This guarantees that all the queries are pertinent, which is a favorable setting with respect to the criterion adopted in this work.

maximum ratio, r_{max}, of non-pertinent with respect to pertinent queries.

For the experiments carried out in this work, both S and S2 were adequately *sampled* in order to produce query sets with increasing ratios, r, of non-pertinent with respect to pertinent queries. To produce a query set with a given ratio r, first all the pertinent queries of the type considered were included in the set and then the remaining queries available for this type were randomly sampled one by one without replacement until the ratio r was reached. Following this procedure, 14, 12 and 15 query sets with increasing r (ranging from 0 to 32, c.f. Sect. 5) were generated for *SINGLE, AND* and *OR* query types, respectively. The ranges of sizes of these sets were: 674−3293, 11,784−925,880, and 992,007−5,395,078, for the *SINGLE, AND* and *OR* query types, respectively.

In page-level KWS experiments, in addition to the number of queries, the total number of *query events*, that is, the number of pairs composed of an image and a query, is also informative. A query event is *pertinent* if the page image is relevant for the query (i.e., the query is actually written in the image). Table 2 also shows the event-level information for the different query types. It is worth noting that the maximum proportions, r_{max}, of non-pertinent with respect to pertinent *query events* are much larger in this case than when measured just in terms of plain queries.

Clearly, spotting *non-pertinent* queries is challenging, since the system may erroneously find other similar queries, which may lead to important precision degradation. Overall, the selected queries constitute rather challenging sets.

4.3 KWS evaluation measures

To assess KWS effectiveness, we employed the standard *recall* and *interpolated precision* measures, which are functions of a threshold used to decide whether a relevance probability $P(R \mid \mathbf{x}, v)$ (see Eq. (2)) is high enough to assume that a word v is in the page $\mathbf{x}$. Interpolated precision is widely used to avoid cases in which plain precision can be ill-defined [16]. Moreover, the popular scalar measure called *average precision* (AP) [22, 33] and the so-called *R-precision* (RP) are also used. The AP is defined as the area under the Recall-Precision curve. On the other hand, the most simple RP measure is defined as the precision (or recall) for some not null threshold such that *recall* is equal to *interpolated precision*. In addition, the maximum value of the *harmonic mean* of precision and recall, called F_1*-measure*, is used also to assess the overall behavior of a search and retrieval system.

4.4 System set-up

In order to build the page-level index, transcribed line images of the training partition were used to train both the optical, and language models.

In this work, hidden Markov models (HMMs) are used for optical modeling. 86 left-to-right character HMMs were trained from the line images, represented as 24-dimensional feature vector sequences computed according to [14]. HMM training consisted in 20 iterations of the Embedded Baum-Welch algorithm [12, 32], followed by 10 iterations of Lattice-Based Extended Baum-Welch Discriminative Training, as described in [31]. Likewise, for better lexicon and language model training, an improved text tokenization was applied which rules white-space among words, punctuation marks and digits (see Table 1 and [25]). A 2-gram word language model was trained using the Kneser-Ney back-off smoothing technique [13]. Meta-parameters associated with 2-gram and HMM training (*grammar scale factor*, *word insertion penalty*, *number of states* per HMM and *number of Gaussians* per state) were tuned using the validation partition. See [25] for more details about these settings.

Finally, using the previously trained models, page-level posterior probabilities of single-word queries, $P(R \mid \mathbf{x}, v)$, were obtained as in Eq. (2) (see Sect. 2), as well as the corresponding probabilities for *AND* and *OR* word-pair queries, according to Eqs. (3–4).

While HMM optical modeling is adopted in this work, it is worth noting that the proposed probabilistic KWS methods, both for single-word and multi-word Boolean queries, can easily be implemented on top of any kind of character-level optical modeling approach. In future work, we plan to test the impact of better optical modeling using Convolutional/Recurrent Neural Networks, as in [3].

5 Results

As shown in Table 2, the maximum ratios (r_{max}) between non-pertinent and pertinent queries are all larger than 1. This ratio is specially large for *AND* queries. In the experiments presented in [18], the whole query sets of Table 2 were used to measure and compare KWS performance for the different query types. This was notoriously unfair for *AND* queries, because the vast majority of them (99.78%) were non-pertinent. In [18], this led to significantly lower KWS performance for *AND* queries which in turn misled to the conclusion that *AND* queries were somehow more "difficult" than *SINGLE* and *OR* queries.

As mentioned in Sect. 1, in this work, we present new empirical results using adequately balanced query sets which allow us to fairly compare the KWS performance of the different query types. To this end, query sets of increasing ratio,

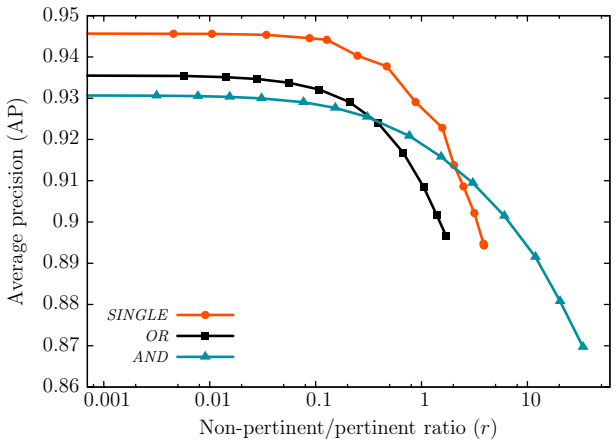

Fig. 2 *Average Precision* (AP) as a function of the ratio between non-pertinent and pertinent queries (r)

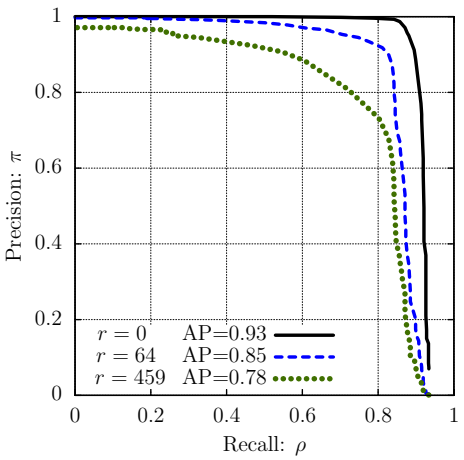

Fig. 3 *Recall-Precision* curves and AP results for word-pair *AND* query sets with extreme values of r

r, of non-pertinent queries were generated for the each type of query, as explained in Sect. 4.2. This ratio was varied in a wide range in order to accurately measure the negative impact on KWS performance of including increasing amounts of non-pertinent queries. The results of this study are shown in Fig. 2, which plots the average precision (AP) as a function of r, for the three types of queries considered.

For r lower than 0.5, the three query types achieve similarly good performance, with AP values greater than 0.92. For larger values of r, AP tends to degrade rather rapidly for all query types, but somewhat less for *AND* queries. It is important to understand that, from a practical point of view, a ratio such as $r = 1$ is already quite large: it would correspond to the unlikely use of an information retrieval system where every other query would be issued to hopelessly try to find information which can *not* actually be found in the indexed collection.

In Fig. 2 both the *SINGLE* and the *OR* curves appear to end prematurely, but this is just because r has reached the maximum possible values, $r_{\max}$, for these types of queries in the relatively small test set used in the present experiments (see Table 2). In contrast, the *AND* curve can still go further down, since $r_{\max}$ in this case is very much larger (459), due to the huge amount of non-pertinent *AND* queries which are possible from the word-pair set described in Table 2. Consequently, only for *AND* queries the degradation of AP when the amount of non-pertinent queries is aggressively increased can be studied. Results of this study are presented in Fig. 3. It shows *Recall – Precision* (R-P) curves and the corresponding AP values for two extreme ratios of non-pertinent queries, namely $r = 459$ (already reported in [18]) and $r = 0$, along with the corresponding results for an intermediate, albeit very large ratio, $r = 64$.

It is fairly clear that most of the degradation is due to false positives produced when searching for non-pertinent word-pairs. Obviously, when trying to find a word-pair which does not actually exist in any of the test images, a perfect system should not produce any spot, unless the confidence threshold is set to 0. But a real system may spot, with non-negligible confidence, images containing words different but similar to those stated in the query. This typically tends to result in degradations of the spotting precision. It is thus gratifying to observe that, even in the most extreme case where the vast majority of queries are non-pertinent, the proposed multi-word KWS approach still provides a decent, usable precision-recall performance (AP = 0.78).

To finish this section, Table 3 reports overall KWS performance in terms of AP, *R-precision* (RP) and maximum F_1*-measure* (F_1^*) figures for low and moderate proportions of non-pertinent queries, $r = 0$ and $r = 1$. We can observe that the three query types behave very similarly, with good

Table 3 Average Precision (AP), R-Precision (RP) and maximum F_1-measure (F_1^*) for the three query sets considered and for r equal to 0 and 1

	$r = 0$			$r = 1$		
	AP	RP	F_1^*	AP	RP	F_1^*
Single	0.946	0.915	0.933	0.927	0.907	0.909
OR	0.935	0.911	0.918	0.909	0.881	0.884
AND	0.931	0.913	0.929	0.919	0.899	0.920

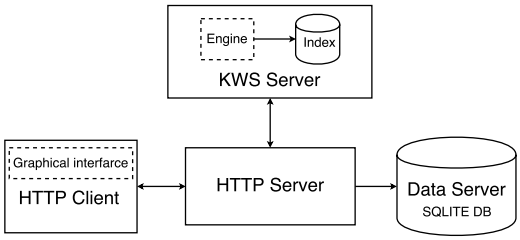

Fig. 4 KWS web demonstrator architecture

comparable performance for each r, and best performance obviously achieved in all the cases for $r = 0$.

It should be finally remarked that the performance achieved in all the cases, even the most adverse ones, is very good, as compared with results reported in recent work on segmentation-free single-word KWS. Moreover, many of these works are based on query sets extracted from the test data which, as discussed throughout this paper, ensures that all the queries are pertinent (i.e., $r = 0$), typically helping to increase the KWS performance. Although not fully comparable with the current work (see comments in Sect. 4.1), the best result obtained in the ICFHR'14 KWS competition with the Bentham dataset only achieved a *mean Average Precision* (mAP)[5] of 0.42 for query-by-example (QbE) KWS. This result was later improved in [28], where a QbE KWS mAP of 0.72 was achieved. The same paper also reported a mAP of 0.86 for QbS KWS, which is more directly comparable with the results of the present paper. Likewise, the winner of the ICDAR'15 KWS competition on the Bentham dataset, a system based on RNN, achieved a mAP of 0.87. Even though RNNs are the current state-of-the-art for HTR optical modeling, it is worth noting that the KWS performance obtained in the present work, based HMM optical models, is advantageously comparable with the best result of the ICDAR'15 KWS competition.

The high degree of usability of the results here presented can be witnessed first hand through real tests using the public demonstration system described in the following section.

6 Demonstration system

In order to provide a user-friendly interface that allows for public testing of the proposed multi-word KWS approach, a demonstrator[6] was implemented with the client-server architecture shown in Fig. 4. It is composed of 4 different modules: *KWS Server*, *HTTP Server*, *Data Server* and *HTTP Client*.

The *KWS Server* module provides single-word confidence scores by looking up an inverted index, which is hierarchically organized in several levels: collection, book and page. It also implements the page-image level Boolean multi-word query logic and probability computations proposed in this work, along with a basic parser which understands the query-string syntax. The *HTTP Server* module is responsible of honoring normal requests from web clients and dynamically builds responses using the data obtained from the *KWS Server* and *Data Server*; that is, word location coordinates and corresponding document information, along with the handwritten text images to be displayed by the client. The *Data Server* is a database which provides the required information about indexed documents: title, description, chapters information, pages information, page images and line bounding boxes, etc. Finally, the *HTTP Client* module implements the GUI. It is thus in charge of interacting with the users, allowing them to pose and edit the query strings, to send these query requests to the *HTTP Server* and to display the query results, namely the retrieved images and the bounding boxes of the spotted keywords.

Some technical specifications of the implemented client-server architecture shown in Fig. 4, are listed below:

- KWS Server
  - Implements a RESTful API using HTTP
  - KWS searches are resolved using the KWS index
  - Each keyword maps to a subindex of items with their confidence score.
- HTTP Server
  - Resolves client requests and dynamically builds responses using PHP
  - Connects to the Data Server to obtain images, book titles, etc (SQL queries)
  - Connects to the KWS Server to bypass the KWS query sent by the user and process the results (HTTP REST petitions).
- HTTP Client
  - Sends standard HTTP requests to the HTTP Server
  - Web pages built using HTML and Javascript on the client side

[5] In general terms, mAP is quite correlated with AP for measuring KWS performance. The use of mAP requires that all the queries are pertinent (see Sect. 4.2 for details).

[6] http://transcriptorium.eu/demots/kws/index.php

Fig. 5 KWS GUI displaying search results for the query "`(jail || prison) (clean || easy)`" at the book level

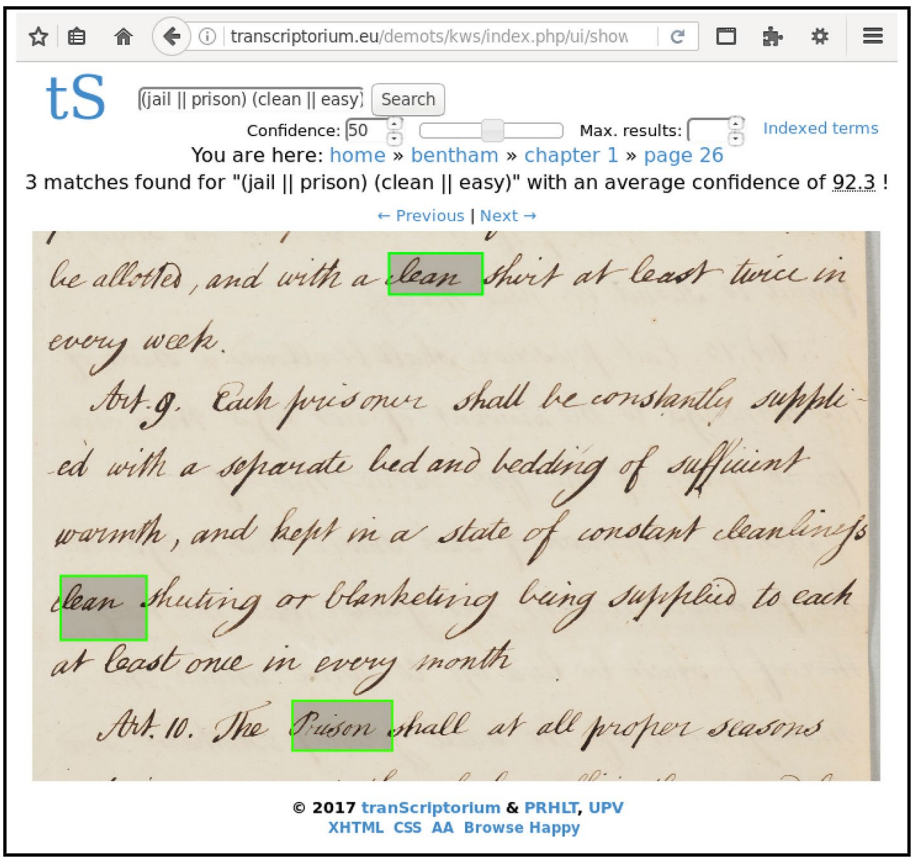

Fig. 6 KWS GUI displaying search results for the query "`(jail || prison) (clean || easy)`" at the page level

- Data Server
  - Stores all the static information from books: title, description, chapters information, pages information, page images and lines bounding boxes.

Boolean operators are represented by the characters "`&&`" or blank, "`||`" and "–" for the operators *AND*, *OR* and *NOT*, respectively. In addition, parenthesis "(" and ")" can be used to unambiguously group keywords and operators. Figures 5 and 6 show search results at the book and the page levels, respectively, corresponding to the query string "`(jail || prison)&& (clean || easy)`".

7 Remarks, conclusion and future work

Following the line-level, single-keyword, probabilistic KWS approach introduced in [27, 29], in this paper we have presented simple but probabilistically consistent approximations to deal, at the page-image level, with queries consisting in Boolean combinations of single-keywords. We have also presented a study to evaluate the search performance of multi-keyword spotting based on these approximations.

The good results achieved support the interest of the proposed methods. Based on these methods a web-based demonstration system has been developed and details of this system are also presented in this paper.

A possible drawback of the KWS approach presented here is that it relies on a predefined lexicon, fixed in the training phase, and therefore, it does not support queries involving out-of-vocabulary keywords. To overcome this limitation, a KWS approach relying on character lattices rather than on word-lattices (see Sect. 2) can be used to compute the required line-level frame word posteriors for character strings which are likely to be real words. This idea has been very successfully used in [4] to index the iconic French Chancery Collection, containing 80,000 images of densely handwritten text in medieval French and Latin.

It is important to remark that the probabilistic multi-word spotting framework formulated in Sect. 3 can be straightforwardly applied without any change to lexicon-free probabilistic indices. In fact, the lexicon-free system developed in [4] does fully support multi-word Boolean *AND / OR / NOT* queries and can be tried online at http://prhlt-kws.prhlt.upv.es/himanis.

In future works, we plan to extend the empirical work by studying the performance achieved for queries entailing more than two keywords and more complex Boolean expressions, including a variety of combinations of *OR*, *AND* and *NOT* operations. In that study, we will also explore how the occurrence frequencies of training and testing words affect the search performance of multi-word queries. As commented in Sect. 4.4, one of our immediate plans is to test the impact of better optical modeling using Convolutional/Recurrent Neural Networks, as in [4], on the proposed probabilistic keyword search methods.

Acknowledgements This work was partially supported by the Generalitat Valenciana under the Prometeo/2009/014 Project Grant ALMAMATER, Spanish MEC under Grant FPU13/06281, and through the EU projects: HIMANIS (JPICH programme, Spanish grant Ref. PCIN-2015-068) and READ (Horizon-2020 programme, Grant Ref. 674943).

References

1. Andreu Sanchez J, Romero V, Toselli A, Vidal E (2014) ICFHR2014 competition on handwritten text recognition on transcriptorium datasets (HTRtS). In: 14th International conference on frontiers in handwriting recognition (ICFHR), 2014, pp 785–790
2. Bazzi I, Schwartz R, Makhoul J (1999) An omnifont open-vocabulary OCR system for English and Arabic. IEEE Trans Pattern Anal Mach Intell 21(6):495–504
3. Bluche T, Hamel S, Kermorvant C, Puigcerver J, Stutzmann D, Toselli AH, Vidal E (2017) Preparatory KWS experiments for large-scale indexing of a vast medieval manuscript collection in the hIMANIS Project. In: 14th International conference on document analysis and recognition (ICDAR). (Accepted)
4. Bluche T, Hamel S, Kermorvant C, Puigcerver J, Stutzmann D, Toselli AH, Vidal E (2017) Preparatory kws experiments for large-scale indexing of a vast medieval manuscript collection in the himanis project. In: 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol. 01, pp 311–316. https://doi.org/10.1109/ICDAR.2017.59
5. Boole G (1854) An investigation of the laws of thought on which are founded the mathematical theories of logic and probabilities. Macmillan, New York
6. Causer T, Wallace V (2012) Building a volunteer community: results and findings from Transcribe Bentham. Digital Humanities Quarterly 6
7. España-Boquera S, Castro-Bleda MJ, Gorbe-Moya J, Zamora-Martinez F (2011) Improving offline handwritten text recognition with hybrid hmm/ann models. IEEE Trans Pattern Anal Mach Intell 33(4):767–779. https://doi.org/10.1109/TPAMI.2010.141
8. Fischer A, Wuthrich M, Liwicki M, Frinken V, Bunke H, Viehhauser G, Stolz M (2009) Automatic transcription of handwritten medieval documents. In: 15th International conference on virtual systems and multimedia, 2009. VSMM '09, pp 137–142. https://doi.org/10.1109/VSMM.2009.26
9. Fréchet M (1935) Généralisations du théorème des probabilités totales. Seminarjum Matematyczne
10. Fréchet M (1951) Sur les tableaux de corrélation dont les marges sont données. Ann Univ Lyon 3^e ser Sci Sect A 14:53–77
11. Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J (2009) A novel connectionist system for unconstrained handwriting recognition. IEEE Trans Pattern Anal Mach Intell 31(5):855–868
12. Jelinek F (1998) Statistical methods for speech recognition. MIT Press, Cambridge
13. Kneser R, Ney H (1995) Improved backing-off for N-gram language modeling. In: International conference on acoustics, speech and signal processing (ICASSP '95), IEEE Computer Society, Los Alamitos, vol. 1, pp. 181–184, https://doi.org/10.1109/ICASSP.1995.479394

14. Kozielski M, Forster J, Ney H (2012) Moment-based image normalization for handwritten text recognition. In: Proceedings of the 2012 international conference on frontiers in handwriting recognition, ICFHR '12, pp 256–261. IEEE Computer Society, Washington. https://doi.org/10.1109/ICFHR.2012.236
15. Lavrenko V, Rath TM, Manmatha R (2004) Holistic word recognition for handwritten historical documents. In: First Proceedings of international workshop on document image analysis for libraries, 2004, pp 278–287. https://doi.org/10.1109/DIAL.2004.1263256
16. Manning CD, Raghavan P, Schutze H (2008) Introduction to information retrieval. Cambridge University Press, New York
17. Marti UV, Bunke H (2002) The iam-database: an english sentence database for offline handwriting recognition. Int J Doc Anal Recogn 5:39–46. https://doi.org/10.1007/s100320200071
18. Noya-García E, Toselli AH, Vidal E (2017) Simple and effective multi-word query spotting in handwritten text images, pp 76–84. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-58838-4_9
19. Pratikakis I, Zagoris K, Gatos B, Louloudis G, Stamatopoulos N (2014) ICFHR 2014 competition on handwritten keyword spotting (h-kws 2014). In: 14th International conference on frontiers in handwriting recognition (ICFHR), 2014, pp 814–819
20. Puigcerver J, Toselli AH, Vidal E (2015) Icdar2015 competition on keyword spotting for handwritten documents. In: 13th international conference on document analysis and recognition (ICDAR), 2015, pp 1176–1180
21. Riba P, Almazn J, Forns A, Fernndez-Mota D, Valveny E, Llads J (2014) e-crowds: a mobile platform for browsing and searching in historical demography-related manuscripts. In: 14th International conference on frontiers in handwriting recognition (ICFHR), 2014, pp 228–233. https://doi.org/10.1109/ICFHR.2014.46
22. Robertson S (2008) A new interpretation of average precision. In: Proceedings of the international ACM SIGIR conference on research and development in information retrieval (SIGIR '08), pp 689–690. ACM, New York. https://doi.org/10.1145/1390334.1390453
23. Romero V, Toselli AH, Vidal E (2012) Multimodal interactive handwritten text transcription. Series in machine perception and artificial intelligence (MPAI). World Scientific Publishing, Singapore
24. Sánchez JA, Romero V, Toselli AH, Vidal E (2016) ICFHR2016 competition on handwritten text recognition on the READ dataset. In: 15th International conference on frontiers in handwriting recognition (ICFHR'16), pp 630–635. https://doi.org/10.1109/ICFHR.2016.0120
25. Toselli A, Vidal E (2015) Handwritten text recognition results on the Bentham collection with improved classical N-Gram-HMM methods. In: 3rd International workshop on historical document imaging and processing (HIP15), pp 15–22
26. Toselli AH, Juan A, Keysers D, González J, Salvador I, Ney H, Vidal E, Casacuberta F (2004) Integrated Handwriting Recognition and Interpretation using Finite-State Models. Int J Pattern Recogn Artif Intell 18(4):519–539
27. Toselli AH, Vidal E, Romero V, Frinken V (2016) HMM word graph based keyword spotting in handwritten document images. Inf Sci 370(C):497–518. https://doi.org/10.1016/j.ins.2016.07.063
28. Vidal E, Toselli AH, Puigcerver J (2015) High performance query-by-example keyword spotting using query-by-string techniques. In: Proceedings of 13th ICDAR, pp 741–745
29. Vidal E, Toselli AH, Puigcerver J (2017) Lexicon-based probabilistic keyword spotting in handwritten text images **(to be published)**
30. Vinciarelli A, Bengio S, Bunke H (2004) Off-line recognition of unconstrained handwritten texts using HMMs and statistical language models. IEEE Trans Pattern Anal Mach Intell 26(6):709–720
31. Young S, Evermann G, Gales M, Hain T, Kershaw D (2009) The HTK book: hidden markov models toolkit V3.4. Microsoft Corporation and Cambridge Research Laboratory Ltd, Cambridge
32. Young S, Odell J, Ollason D, Valtchev V, Woodland P (1997) The HTK book: hidden markov models toolkit V2.1. Cambridge Research Laboratory Ltd, Cambridge
33. Zhu M (2004) Recall, precision and average precision. Working paper 2004-09 Department of Statistics and Actuarial Science–University of Waterloo