Probabilistic Indexing and Search for Information Extraction on Handwritten German Parish Records

Eva Lang
Archiv des Bistums Passau (Germany)
eva.lang@bistum-passau.de

Joan Puigcerver, Alejandro Héctor Toselli and Enrique Vidal
PRHLT - Universitat Politècnica de València (Spain)
[joapuipe,ahector,evidal]@prhlt.upv.es

***Abstract*—We endeavor to perform very large scale indexing of an ancient German collection of manuscript parish records. To this end we will compute "probabilistic indexes" (PIs), which are known to allow for very accurate and efficient implementation of (single-)keyword spotting. PIs may become prohibitively large for vast manuscript collections. Therefore we analyze simple index pruning methods to achieve adequate tradeoffs between memory requirements and search performance. We also study how to adequately deal with the large variety of non-ASCII symbols and handwritten word spelling variations (accents, umlauts, etc.) which appear in this kind of historical collections. Finally, and most importantly, since most of the images of the collection we aim to index are handwritten tables, we explore the use of PIs to support structured queries for information extraction from untranscribed handwritten images containing tabular data. Empirical results on a small, but complex and representative dataset extracted from the collection considered confirm the viability and adequateness of the chosen approaches.**

I. Introduction

Libraries, archives and other cultural institutions all over the world are making accessible large amounts of (untranscribed) digitized handwritten documents. This fact is spurring the development of handwriting processing technologies; namely automatic/assisted handwritten text recognition (HTR) and keyword spotting (KWS), to provide access to the textual contents of the images. Examples of the kind of untranscribed documents requiring urgent text access to their contents are birth, marriage, and death records, military draft records, court records, census records, property registers, etc.

Here we deal with a German handwritten parish record collection (C. XVI-XVIII), held by the Passau Diocesan Archives. It features baptism, marriage and death records such as those shown in Fig.1. Aiming at allowing textual search and at extracting information contained in these records, we adopt the probabilistic indexing and search approach already applied to other old, vast document collections such as Chancery [1].

Here, we present empirical work on a relatively small but fully representative dataset extracted from this collection as a preparatory step towards undertaking the whole indexing process. In this work, we have identified and provided adequate solutions to the following two issues: a) a large variety of non-ASCII symbols appear in the images which may be difficult to be typed on modern keyboards; b) slightly different spelling variations of handwritten words (e.g., accents, umlauts, etc.) entail a waste of indexing probability mass which may significantly hinder search performance.

Besides reporting spotting results for single-word queries, an important contribution of this work is to explore the use of PIs to support structured, multiple-word queries aiming at information extraction from records consisting in untranscribed handwritten tables.

II. Handwritten Passau Parish Record Collection

The dataset provided by the Passau Diocesan Archives contains information about the parishioners who were baptized, who married and who died within the geographic boundaries of the various parishes of the Diocese of Passau. The scans originate from more than 100 pastoral districts with their own record keeping, which started in the late 16th century by the order of the Church and is carried out today.

The 289 images of the dataset used in this work (see Sec. II-B) were selected from a subset of 57 222 scans of more than 800 000 sacramental register images[1]. This dataset samples the full collection and demonstrates the evolution of the three record types over time. The images show a great variety in the evolution of handwriting, record keeping and more and more standardized table forms introduced in the early 19th century. Examples of images from this dataset are shown in Fig. 1.[2]

Each record in the register books refers to a sacramental event and therefore provides a reference to the person, who was baptized on a given day, a couple, which was engaged and later married, or the person who died. Besides the names of the individual, references to occupation, locations of living, priests and witnesses or doctors are given. In the baptismal and wedding entries, also the names of the parents are kept and, in death records, the illnesses and reason of death are recorded.

Therefore, the register books are not only of interest to family researchers, who explore their personal history, but also demonstrate the social, economic or humanitarian history. The sacramental records of the Passau Diocesan Archives are accessible through a database[3] developed by the Archives.

A. Transliteration

This (or other) ancient German record collection(s), generally contain large numbers of non-ASCII symbols. In addition,

[1]Openly available at http://data.matricula-online.eu/de/deutschland/passau

[2]The ground-truthed data set is publicly available to download from zenodo: http://doi.org/10.5281/zenodo.1296322

[3]http://gendb.bistum-passau.de

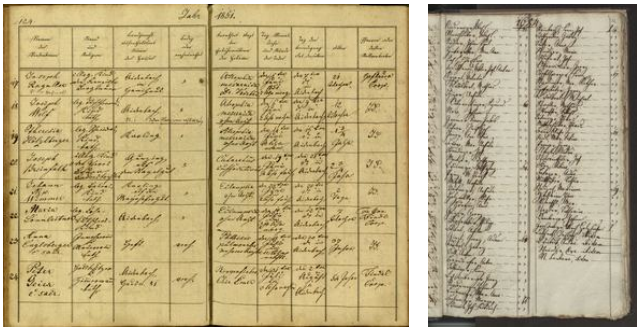

Fig. 1. Some page image examples of the PASSAU dataset.

rather inconsistent slightly different spelling variations of the same word (e.g., accents, umlauts, underdots, tie bar, etc.) appear in the images (and often also in ground truth, reference record transcripts). In this dataset, the character set used for the reference transcripts accounts a total of 263 UTF-8 different symbols, most of which are/contain non-ASCII characters corresponding to archaic symbols used in old German writing. It is worth noting that most of such characters can not be typed on standard keyboards.

A common practice in conventional search engines and key word spotting alike is to retrieve all instances of a given query without matter of spelling variations. To this end, every character or symbol is *transliterated* by *case folding* and, if necessary, by removing diacritics and mapping non-ASCII symbols onto their ASCII equivalents. The benefits are two-fold: a) simplify the composition of queries and b) avoid the waste of probability mass which often leads to degrade search performance. Table I shows examples of applied character transliterations with which the number of different characters in the reference transcripts was reduced to from 263 to 102. The impact of transliteration is studied in Sec, V.

TABLE I
EXAMPLES OF CHARACTER TRANSLITERATION.

Remov. Diacrit.		Non-ASCII to ASCII					
ċ, č, ć	C	Æ, æ	AE	ij	II	ŋ	EN
è, ē, ë	E	Œ, œ	OE	ß	SS	ǫ	US
m̄, m̆, m̨	M	p, ṕ	PRO	đ	DE	ð	DER

B. Experimental Dataset and Training/Test Partition

Table II shows important details of the 289 images dataset used in this work. 179 images were selected for training, 21 for validation and the remaining 89 for testing. All these three partition blocks contain a similar proportion of images from documents dated before and after 18th century in order to cope with the most important variations of handwriting styles and record and table formats.

The "Test" set of images is used for plain, single-word KWS experiments, while "TabTest" is a subset of "Test" containing only tabular data images, generally lying across two contiguous pages as in two of the examples of Fig. 1. It is used to measure the performance of structured multiword queries aimed at information retrieval from table images.

TABLE II
THE PASSAU EXPERIMENTAL DATASET.

	Train+Val	Test	TabTest
Number of pages	200	89	44
Number of lines	29 314	16 376	11 710
Running words	72 848	37 354	21 027
Running words excluding punctuation	–	26 709	15 141
Lexicon size	12 381	6 532	3 455
Number of different characters	220	187	119
Lexicon size after transliteration	11 160	5 801	3 141
Number of transliterated different chars	99	87	73

III. PROBABILISTIC INDEXING AND SEARCH

The probabilistic framework which supports the proposed indexing and search approaches is outlined in this section.

A. From Filler-based KWS to Probabilistic Indexing

The *Filler* approach to KWS is among the most popular and successful techniques for training-based, query-by-string, lexicon-free and segmentation-free KWS. It was first proposed for handwriting KWS in [2], using optical character hidden Markov models. A method based on related ideas was also developed for recurrent (BLSTM) neural network optical character models [3].

In its original form, the Filler approach only produces moderately accurate KWS results. Moreover, it incurs a very high computational cost for the search of each given keyword. The latter shortcoming was much alleviated in [4], by precomputing character lattices (CL) which allow up to 98% reductions in search time, with identical KWS accuracy. Better accuracy was achieved in [5], by incorporating a 2-gram character language model to model lexical-like context, but at the expense of an even larger search time. Later, KWS accuracy was further increased significantly in [6] using 6-gram character language models, while also keeping a low search cost by using the CL-based technique introduced in [4].

In [7] a probabilistic interpretation of the Filler approach was presented which showed that the log-likelihood ratio word confidence score used in the Filler approach is in fact equivalent to a simple, Viterbi approximation to the *relevance probability* (RP) $P(R \mid \mathbf{c}_v, \mathbf{x})$, where R is a binary random variable which models whether an image region $\mathbf{x}$ is *relevant*, or not, for a query keyword v, formed by the concatenation of characters $c_1, c_2 \ldots, c_m \stackrel{\text{def}}{=} \mathbf{c}_v$. The RP concept was introduced and is commonly used in the field of information-retrieval [8] and it was also implicitly adopted in the successful approach to lexicon-based KWS presented in [9]. The developments of [7] also led to an accurate way of computing the true $P(R \mid \mathbf{c}_v, \mathbf{x})$, allowing the new approach to significantly overcome KWS results of traditional Filler approximations. These improvements where consolidated in [10] where, using N-gram character language models, the KWS accuracy was further improved while keeping the low search times of previous methods based on CLs.

However, the methods of [10] still entail significant computation load on the CL of each image-region for each query keyword. Even though it is very much faster than all the

Filler versions not based on CLs, computing time still becomes completely prohibitive for very large text image collections.

To overcome this limitation a (probabilistic) word index could be used, as in the lexicon-based approach of [9]. However, lexicon-free KWS can not rely on any a-priory fixed lexicon because it must support queries consisting of arbitrary character sequences. Of course, an ideal, oracle-based system should only index character sequences corresponding to words which do appear in the image collection to be indexed. Then, if a non-indexed character sequence is queried, the system would just assume it has a null RP.

Obviously, using an oracle is not an option. But from the huge amounts of different, probabilistically scored character sequences contained in the CL of an image region $\mathbf{x}$, one can extract some, or many, sequences which have a non-negligible probability of being actual words written in $\mathbf{x}$. Then these sequences, which we refer to as *"pseudo-words"*, can be easily indexed, along with the corresponding RPs and geometric information (bounding boxes), to support extremely fast search performance. Even though only a finite number of pseudo-words can be indexed, the approach can properly be considered *lexicon-free*: Clearly, any character sequence can be queried, although those which have negligible probability of actually being written in the text images will just get a null RP.

This idea was very successfully used in [1] to index the iconic French Chancery Collection, containing 80 000 images of densely handwritten text in medieval French and Latin. The resulting probabilistic index (PI) supports almost instantaneous full free-text search over the whole collection and is now available for public use at http://prhlt-kws.prhlt.upv.es/himanis.

B. Building the Probabilistic Index

In this work convolutional and recurrent neural networks (CRNN) are used for optical character modeling. Specifically, we employed the HTR Laia Toolkit [11] based on the Torch machine learning platform. Adopted CRNN architecture is commented is Sec. IV-C.

After training, the CRNN computes sequences of character posterior probabilities from a text line image, previously cleaned and contrast-enhanced using a Sauvola's based method [12], and normalized to 64-pixel height maintaining the aspect ratio.

Lexical and linguistic context is explicitly modeled by means of statistical character N-grams, estimated using the training image transcripts[4] and represented as a weighted finite-state transducer. The CRNN output character probabilities, scaled with character priors, are then incorporated to the transducer edges. The required CLs are finally obtained by beam search Viterbi decoding using the Kaldi toolkit [13].

The CL obtained from a given text line image, $\mathbf{x}$, represents a huge number of transcription hypotheses (in the form of character sequences, $\mathbf{c}$), along with their corresponding probabilities and geometric character boundaries. As discussed in [9] (see also [10]), it provides a good approximation to the joint probability distribution $p(\mathbf{c}, \mathbf{x})$.

[4]With the SRILM Toolkit: http://www.speech.sri.com/projects/srilm

Finally, following the CL-based indexing method outlined in Sec. III-A, a set of n-best scored word-like subpaths are extracted from the CL. Such subpaths define the character sequences referred to as "pseudo-words", along with their geometric locations and the corresponding RPs [1]. These three data items are then stored in the resulting PI.

Sec.IV-C provides details about (meta-)parameters adopted in each step: CRNN, CL generation and RP computation.

C. Structured Multi-Word Queries

As shown in [14], a PI provides adequate support for KWS queries consisting of boolean (AND, OR, NOT) combinations of multiple keywords. Here we extend these ideas to support *structured multi-word queries*, aimed at information retrieval in text images containing *tabular data*. More specifically, we explain how to deal with queries of the form "⟨`column-heading, column-content`⟩", where `column-heading` is an AND combination of table heading words and `column-content` is a (single) keyword. For example, the query "⟨`NAMEN VERSTORBENEN, WOLF`⟩" ("⟨`DECEASED NAME, WOLF`⟩" in English), should retrieve the handwritten word "*Wolf*" which appears in the third row of the first column of the top-left image in Fig. 1 (see also Fig. 2 for a zoomed view and additional details).

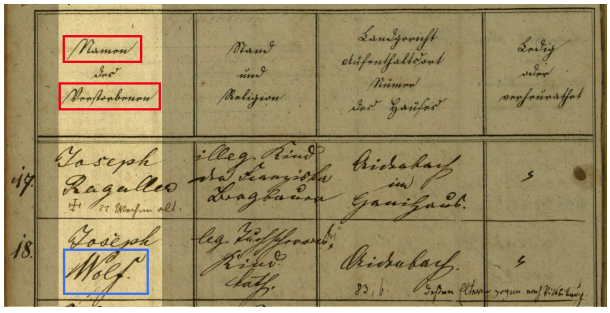

Fig. 2. Geometric reasoning for the column-wise multi-word structured query: "<NAMEN VERSTORBENEN, WOLF>".

The retrieval process is carried out in two steps for each table image. First, column-heading words with RP higher than the given threshold τ are retrieved. Simple geometric restrictions are applied in this step: candidate spots must be close enough to each other and all must loosely be located in upper regions of the image. This is supported by the location information (word bounding boxes) stored in the image PI. Let $\mathbf{h} \stackrel{\text{def}}{=} \{h_1, h_2 \ldots, h_I\}$ be the set of column-heading query words. Abusing the notation, each h_i will be also understood as the boolean event that h_i is relevant for some image region. Let $R(h_i)$ denote the RP of h_i. Note that repeated instances or spots of some of these words may have been retrieved in the image region of a column-heading. Let $s_{i1}, \ldots, s_{iJ_i}$ denote the $J_i \geq 1$ different spots of the query word h_i and let $R(s_{ij}) \stackrel{\text{def}}{=} P(R \mid \mathbf{c}_{h_i}, \mathbf{x}_{ij})$ be the RP of the word h_i in the image location $\mathbf{x}_{ij}$, where $\mathbf{c}_{h_i}$ is the character spelling of the word h_i (cf. Sec. III-A). Then, according to [14], the RP of

the AND combination for the words in $\mathbf{h}$ is computed as:

$$\begin{aligned} R(\mathbf{h}) &= R(h_1 \wedge h_2 \cdots \wedge h_I) \\ &\approx \min_{1\le i\le I} R(h_i) \approx \min_{1\le i\le I} \max_{1\le j\le J_i} R(s_{ij}) \end{aligned} \quad (1)$$

Second, the column-content word v is searched for through the column-wise regions delimited by the horizontal span of the spotted column-heading word bounding boxes (see Fig. 2). Again, only spots with RP higher then τ are retrieved and the search is geometrically supported by the location information stored in the image PI. Let $v_1, \ldots, v_K$, $K \geq 1$, be the different spots of v retrived in column locations $\mathbf{x}_1, \ldots, \mathbf{x}_K$ and let $R(v_k) \stackrel{\text{def}}{=} P(R \mid \mathbf{c}_v, \mathbf{x}_k)$ (cf. Sec. III-A) be the RP of the k-th spot. From the discussion in [14], the RP of the column-content word v in the considered column is computed as:

$$R(v) \approx \max_{1\le k\le K} R(v_k) \quad (2)$$

Finally, again according to [14], the RP of a column-wise structured multi-word query is computed as:

$$R(\langle \mathbf{h}, v \rangle) = R(\mathbf{h} \wedge v) \approx \min\big(R(\mathbf{h}), R(v)\big) \quad (3)$$

Following the same ideas, more complex boolean combinations can be straightforwardly supported. Row-wise and word-sequence combinations can be supported too by using the location information of the spotted words. However, the study of these more complex queries is left for future research.

IV. Experimental Setup

A. Evaluation Measures

The standard *recall* and *interpolated precision* measures [8] are used here. Results are reported in terms of both global and mean *average precision* (AP and mAP, respectively). While the AP is computed from a *global* ranked list containing all the results from all queries, the mAP is the mean of the APs of the isolated queries. Both are standard quality measures of ranking, but the AP also evaluates consistency of the scores across multiple queries. They are sometimes known as micro and macro AP [15]. Moreover, we will report the maximum recall achieved with at least 10% precision (MxRc_{10}).

Finally, we also evaluate naive KWS based on 1-best HTR transcripts. Since no scores are available, the RP curve would degenerate to a single point. Yet, the interpolated precision allows computing both AP and mAP.

B. Query Sets

Several criteria can be adopted to define a set of keywords to be used for KWS evaluation. In this work we adopt one of the most common criteria, where most of the words seen in the test set are selected as keywords. Besides being a meaningful choice from an application point of view, it ensures that all the keywords are relevant (appear in one or more test images), thereby allowing mAP to be properly computed.

For single-keyword query spotting, all the words longer than 1 character of the test-set lexicon are used, making a total of 5 725 transliterated query words. On the other hand, for multi-word queries we asked a user to provide application-meaningful queries of the form "⟨column-heading, column-content⟩". The user aimed these queries at obtaining information registered in some of the 44 test-set images which contain tabular data. A set of 363 queries of this type was collected. The number of different words in the column-heading parts of these structured multi-word queries range from 1 to 6, while all the column-content parts contain just 1 word. Examples of these queries are: ⟨ORT, STEINERLEINBACH⟩, ⟨TAUF TAG, APRIL⟩ ⟨KRANKHEIT ARZT, FRAISEN⟩, ⟨NAMEN DES BRAEUTIGAMS, JOSEF⟩, ⟨NAMEN DER BRAUT, MARIA⟩, ⟨TAG MONAT JAHR TODES, 1879⟩, etc.

C. System Setup

The CRNN architecture adopted here consists of four convolutional layers followed by three recurrent bidirectional long short-term memory layers. Regarding layer parameter setting (dropout, maxpooling, activation function, etc.), we employed the default ones of the HTR Laia Toolkit [11], with exception of the ones corresponding to number of convolutional features: 12, 24, 48 and 48, and the ones of convolutional kernel sizes: 7×7, 5×5, 3×3 and 3×3. A softmax output layer computes the probabilities of each character in the training alphabet plus a non-character symbol.

The CRNN was trained with the RMSProp method [16] on minibatches of 32 examples, using a base learning rate of 0.0005, to minimize the CTC cost function [17]. We stopped the optimization procedure when the error on the development set did not decrease for 50 epochs. Two optical character sets were trained (see Tab. II): one of 220 characters using the original transcripts and another of 99 characters using transliterated transcripts. These sets were used in the experiments referred to as *late* and *early transliteration*, respectively.

N-gram character language models were estimated (with Kneser-Ney back-off smoothing) [18]) from the original and from the transliterated transcripts. Depending where transliteration is applied, we distinguish between: *early transliteration* and *late transliteration*. The latter is applied directly to CLs, while the former is applied from the beginning by training the CRNN using transliterated transcripts. According to this distinction, four N-gram character language models were estimated: A 6-gram model was used for late transliteration, while 0-gram, 3-gram and 6-gram models were used in the early transliteration experiments.

The trained CRNN, along with each character language model, were subsequently used with the Kaldi decoder to produce CLs for all the test-set image lines. One decoding pass with beam set to 15 was carried out in all the cases. In the late transliteration experiments, the character sequences associated with the edges of the CLs obtained using the original character set, were transliterated as described in Sec. II-A.

Using the CLs produced (and transliterated when appropriate) as described above, diacritic-free and case-folded PIs were obtained as explained in Sec. III-B.

V. KWS Performance Results

Two sets of experiments were carried out. The first one is intended to evaluate single-keyword KWS performance for early and late transliteration and different character N-gram orders. The second one focuses on assessing search performance for structured multi-word queries aimed at information extraction.

A. Single-Keyword Queries

Table III reports AP, mAP and MxRc_{10} results obtained for CLs produced with early and late transliteration and different N-gram orders. In addition, results are also reported for "degenerate lattices" consisting of a single, linear path with the plain 1-best hypothesis of an HTR recognizer using the same optical and 6-gram language models as those used for KWS. This is equivalent to naive KWS based on plain-text keyword searching in the single-best HTR transcripts. The character error rates of these transcripts were 16.6% and 20.9% for the early and late transliterations, respectively.

TABLE III
KWS PERFORMANCE FOR SINGLE-WORD (PLAIN) AND STRUCTURED MULTI-WORD (STRUCT) QUERIES, FOR EARLY/LATE TRANSLITERATION AND CLS GENERATED WITH DIFFERENT N-GRAMS ORDERS (LM).

	Transliteration	Latt-type	Char LM	AP	mAP	MxRc_{10}
PLAIN	Early	CLs	none	0.701	0.661	0.861
	Early	CLs	3-gram	0.712	0.677	0.876
	Early	**CLs**	**6-gram**	**0.746**	**0.692**	**0.886**
	Late	CLs	6-gram	0.692	0.662	0.854
	Early	1-best	6-gram	0.559	0.387	0.680
	Late	1-best	6-gram	0.492	0.331	0.613
STRUCT	**Early**	**CLs**	**6-gram**	**0.905**	**0.921**	**0.955**

Figs. 3 shows interpolated recall-precision (R-P) curves corresponding to early and late transliteration, along with a single R-P point, corresponding to naive KWS based on 1-best recognition hypotheses obtained with early transliteration. A character 6-gram language model was used in all the cases.

As observed, all the results improve with increasing N-gram order, assessing the importance of leveraging linguistic context. The benefits of transliterating right from the beginning of the indexing process are also clear from the results. Finally, results confirm that only relatively poor performance can be achieved through naive KWS based on 1-best HTR transcripts.

B. Structured Multi-Word Queries for Information Retrieval

AP, mAP, MxRc_{10} results for structured multi-word queries aimed at information retrieval in tabular text images are reported in the last row of Tab. III and in Fig. 3.

Retrieval performance is clearly much better in this case than in the case of single-keyword queries. Note that the single-keyword query set encompassed all the test-set words, including function words and many other rather "uninteresting" words. In contrast, the structured queries used in this experiment are real queries issued by a user aiming to retrieve real data from the text images. Therefore, while the conditions are not properly comparable, we think the results obtained with natural queries should be considered more realistic.

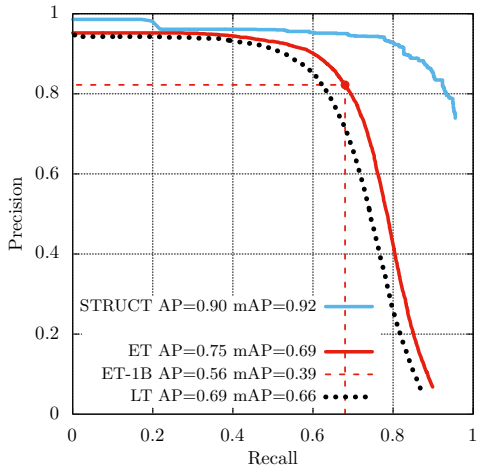

Fig. 3. *Recall-precision* (R-P) curves for single-keyword KWS using CLs obtained with early (ET) and late (LT) transliteration. For ET, a single R-P point (1B-6g) is also shown for naive KWS based on 1-best recognition hypotheses. The R-P curve for ET structured multi-word queries (STRUCT) is also shown. Character 6-gram models were used in all the curves.

Note as well that the very structured nature of these queries might have also contributed to achieve better performance: For this kind of queries, several words must be successfully spotted in an image and geometric constraints further limit the set of final candidate spots. This improves precision because many possible false alarms just fail to survive the underlying relational and geometric conditions.

VI. Probabilistic Index Trimming

As previously commented, for vast manuscript collections, PIs may become huge and require prohibitively large amounts of storage. On the other hand, because of their *lexicon-free indexing* construction, PIs contain large quantities of pseudo-words which probably will never be used in any real query. It is important to understand that the (large number of) possibly useless pseudo-word spots do not harm precision-recall performance, but do result in large storage overheads. However, it is this very overhead what grants the flexibility of fast searching for arbitrary character strings, rather than specific words (as in *lexicon-based* KWS).

Ideally one would like to get rid of useless pseudo-words and retain only character strings which are likely useful for information search; but this is like finding a few needles in a big haystack. Fortunately, it can be observed that most indexed "rare" pseudo-words have very low RP. Therefore we can try to use the RP to prune out those spots which do not lead to a significant negative impact on search performance.

Fig. 4 plots the evolution of single-keyword AP and mAP for increasing index size reduction, obtained by increasing a pruning RP threshold (RPT). It can be seen that AP remains practically invariant to reductions up to about 90%, while mAP starts to suffer some degradations when more than 40% of the index spots (those with RP$< 10^6$) are pruned out.

Table IV reports PI sizes per page image for conservative and more aggressive RP thresholds (those shown in Fig. 4), along with the corresponding relative AP and mAP degradations. The actual amount of storage needed per indexed page image is also reported, along with the ratio of this

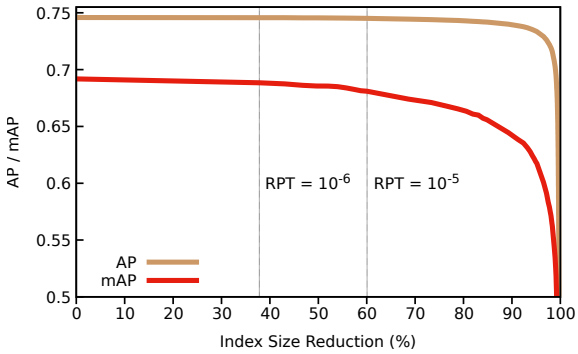

Fig. 4. Probabilistic Index Pruning: AP and mAP for increasing index size reduction (in %), obtained by pruning out indexed spots whose RP is lower than an increasing RP threshold (RPT).

size with respect to the actual average size of the indexed (JPEG) images. For comparison, two extreme cases in terms of performance degradation are reported: indexing using 1-best HTR transcripts and indexing just a hypothetically known exact lexicon of the collection. The index size is greatly reduced in both cases, but the 1-best index is largely useless and the lexicon-based one is only an hypothetical lower bound.

TABLE IV
PROBABILISTIC INDEX MEMORY USAGE AND SEARCH ACCURACY TRADEOFFS. RATIO IS THE QUOTIENT BETWEEN INDEX SIZE AND AVERAGE INDEXED IMAGE SIZE (1 823 KB IN THE PASSAU DATA SET).

	Spots/page	KB/page	Ratio	ΔAP	ΔmAP
No trimming	56 953	897	0.490	0.000	0.000
Conservative trimming	35 412	552	0.304	0.000	-0.003
Aggressive trimming	22 742	356	0.195	-0.001	-0.011
Naive 1-best indexing	296	5	0.003	-0.187	-0.305
Exact lexicon oracle	2 853	46	0.025	0.000	0.000

VII. CONCLUSIONS

Highly efficient lexicon-free single-keyword KWS has been shown to be very well supported by Probabilistic Indices. It has been also shown how these indices can be easily used to very effectively support structured queries involving many words, which allow for complex information retrieval in text images containing tabular data. The techniques used to obtain the proposed probabilistic indices have been outlined and evaluated. The empirical work presented constitutes a preparatory step before undertaking the actual indexing of a very large collection containing hundreds of thousands of images of historical handwritten registers.

A real demonstrator of the indexing and search techniques developed and evaluated in this work is publicly available at http://transcriptorium.eu/demots/kws-Passau (it allows single-word and multiple-word boolean queries, but the structured tabular queries introduced in this paper are not yet supported).

VIII. ACKNOWLEDGMENTS

This work was partially supported by the Spanish MEC under FPU grant FPU13/06281, by the Generalitat Valenciana under the Prometeo/2009/014 project grant ALMAMATER, and through the EU project READ (Horizon-2020 programme, grant Ref. 674943).

REFERENCES

[1] T. Bluche, S. Hamel, C. Kermorvant, J. Puigcerver, D. Stutzmann, A. H. Toselli, and E. Vidal, "Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project," in *2017 14th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR)*, vol. 01, Nov 2017, pp. 311–316.

[2] 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.

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

[4] A. H. Toselli and E. Vidal, "Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents," in *Proc. of the 12th Int. Conf. on Document Analysis and Recognition (ICDAR '13)*. Washington, DC, USA: IEEE Computer Society, 2013.

[5] A. Fischer, V. Frinken, H. Bunke, and C. Suen, "Improving HMM-Based Keyword Spotting with Character Language Models," in *Doc. Anal. and Recog. (ICDAR), 2013 12th Int. Conf. on*, Aug 2013, pp. 506–510.

[6] A. H. Toselli, J. Puigcerver, and E. Vidal, "Context-aware lattice based filler approach for key word spotting in handwritten documents," in *Document Analysis and Recognition (ICDAR), 2015 13th Int. Conf. on*, Aug 2015, pp. 736–740.

[7] J. Puigcerver, E. Vidal, and A. H. Toselli, "Probabilistic interpretation and improvements to the HMM-filler for handwritten keyword spotting," in *2015 13th Int. Conf. on Document Analysis and Recognition (ICDAR)*, Aug 2015, pp. 731–735.

[8] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[9] A. H. Toselli, E. Vidal, V. Romero, and V. Frinken, "HMM Word Graph based Keyword Spotting in Handwritten Document Images," *Information Sciences*, vol. 370-371, pp. 497–518, 2016.

[10] A. H. Toselli, J. Puigcerver, and E. Vidal, "Two methods to improve confidence scores for lexicon-free word spotting in handwritten text," in *2016 15th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR)*, Oct 2016, pp. 349–354.

[11] J. Puigcerver, "Are multidimensional recurrent layers really necessary for handwritten text recognition?" in *2017 14th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR)*, vol. 01, Nov 2017, pp. 67–72.

[12] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, pp. 225–236, 2000.

[13] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[14] E. Noya-García, A. H. Toselli, and E. Vidal, *Simple and Effective Multi-word Query Spotting in Handwritten Text Images*. Springer International Publishing, 2017, pp. 76–84.

[15] F. Perronnin, Y. Liu, and J. M. Renders, "A family of contextual measures of similarity between distributions with application to image retrieval," in *2009 IEEE Conf. on Computer Vision and Pattern Recognition*, June 2009, pp. 2358–2365.

[16] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, 2012.

[17] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. of the 23rd Int. Conf. on Machine Learning*, ser. ICML '06. NY, USA: ACM, 2006, pp. 369–376.

[18] R. Kneser and H. Ney, "Improved backing-off for N-gram language modeling," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP '95)*, vol. 1, CA, USA, 1995, pp. 181–184.