2017 14th IAPR International Conference on Document Analysis and Recognition

Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project

Théodore Bluche*, Sebastien Hamel†, Christopher Kermorvant‡, Joan Puigcerver§,
Dominique Stutzmann†, Alejandro H. Toselli§ and Enrique Vidal§

*A2iA SAS
Paris, France
Email: tb@a2ia.com

†IRHT - CNRS
Paris, France
Email: sebatien.hamel@irht.cnrs.fr
dominique.stutzmann@irht.cnrs.fr

‡TEKLIA SAS
Paris, France
Email: kermorvant@teklia.com

§PRHLT - UPV
Valencia, Spain
Email: joapuipe@upv.es
ahector@prhlt.upv.es
evidal@prhlt.upv.es

***Abstract*—Making large-scale collections of digitized historical documents searchable is being earnestly demanded by many archives and libraries. Probabilistically indexing the text images of these collections by means of keyword spotting techniques is currently seen as perhaps the only feasible approach to meet this demand. A vast medieval manuscript collection, written in both Latin and French, called *"Chancery"*, is currently being considered for indexing at large. In addition to its bilingual nature, one of the major difficulties of this collection is the very high rate of abbreviated words which, on the other hand, are completely expanded in the ground truth transcripts available. In preparation to undertake full indexing of *Chancery*, experiments have been carried out on a relatively small but fully representative subset of this collection. To this end, a keyword spotting approach has been adopted which computes word relevance probabilities using character lattices produced by a recurrent neural network and a N-gram character language model. Results confirm the viability of the chosen approach for the large-scale indexing aimed at and show the ability of the proposed modeling and training approaches to properly deal with the abbreviation difficulties mentioned.**

I. Introduction

With the growth of high-capacity storage devices and faster large-scale digitizing systems, more and more datasets of digitized document collections are being made available by archives and libraries. To make such large image datasets useful, indexing techniques which can quickly and accurately extract textual information from untranscribed digital images are highly desired for. In the literature, several approaches are proposed for indexing handwritten documents, most of them based on *keyword spotting* (KWS) methods.

The most effective KWS approaches generally require training data to estimate the parameters of their statistical models, like hidden Markov models (HMMs) [1], [2] and/or the weights of their neural network units [3], [4]. When a large collection of document images is considered, it is very common to have moderate amounts of transcribed images available – or if they are not, the cost of producing these transcripts is often negligible with respect to the overall cost of the indexing project. These transcribed images can serve as training data for such training-based KWS approaches.

Specifically in this case we consider the use of a training-based, segmentation-free *query-by-string* (QbS) KWS approach proposed in [5]. It relies on *character lattices* (CLs) produced by a holistic, character-N-gram driven, handwritten text recognition (HTR) system. An important change introduced in the present work is that character-level optical likelihoods are now obtained by means *recurrent neural networks* (RNN), rather than character HMMs as in [5]. We aim to use this approach to cope with the main challenge set out in the European project HIMANIS, namely, very large scale indexing of a vast medieval manuscript collection, handwritten in both Latin and French, called *"Chancery*".

A mandatory preparatory step for any large-scale indexing project is to perform comprehensive experiments to validate the techniques adopted and assess the the quality of indices produced with different parameter settings, such as the order of N-gram models, etc. To this end, we have used a relatively small, but still fully representative set of around 400 transcribed text image regions, extracted from the full collection.

Taking into account the intended application, evaluation is not (mainly) based on the geometric accuracy of the spotted words, as it is usual in most KWS papers and competitions. Instead, larger image regions such as lines, or even paragraphs, are considered as search targets. This is in line with the goals pursued in other current important projects such as VTM[1] and READ[2], as well as in many recent papers on KWS (see [6], [7], [4], [2], [8], e.g.).

The medieval documents considered here, like those in many other medieval collections, entail two important linguistic challenges. Namely, they are written in more than one language and they are heavily abbreviated, specially the text parts written in Latin. The latter problem is particularly insidious because the (only) image transcripts generally available are "modernized" versions where all the abbreviations are completely expanded. As discussed in [9], this constitutes a serious drawback to train adequate optical character models (HMMs in particular). Moreover, according to the requirements of the search functionality aimed at in HIMANIS, textual queries must allow (only) modernized word forms.

[1] http://vtm.epfl.ch.
[2] https://read.transkribus.eu.

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.59

However, as the results of the present work show, all the indexing challenges, including those entailed by the multilingual and abbreviated texts, are very successfully approached by the proposed combination of RNN optical character models [10], character N-grams, and CL-based KWS [5].

The rest of the paper is organized as follows. The next section provides background of the RNN-based HTR system adopted and essential details of the character lattices produced. Sec. III overviews the KWS statistical framework and the specific approach here proposed. Sec. IV describes in general the manuscript collection targeted for large-scale indexing and the representative dataset selected for the preparatory experiments. This section includes also information about dataset partitions, selection of query sets, evaluation measures and experimental setup. Results and discussion are given in Sec. V. Finally, Sec. VI summarizes the work presented and draws conclusions.

II. Neural Network Recognition and Lattice Production

Our indexing approach is based on a *"rich decoding"* of each text line image, which yields a *character lattice* (CL) as a result. CLs are produced by a handwritten text line recognition system comprising a deep recurrent neural network (RNN) and a statistical character language model.

A. Handwritten Text Recognition

The RNN predicts sequences of character probabilities from the gray level text line image. It is made of eight convolutional layers followed by two recurrent long short-term memory layers. A softmax output layer predicts the probabilities of 104 characters and a blank symbol. The architecture of the network is illustrated in Fig. 1, and described in details in [10]. The network is trained by gradient descent with the RMSProp method [11] and the Connectionist Temporal Classification (CTC [12]) objective. The language models are statistical character N-grams estimated on the image transcripts using IRSTLM [13].

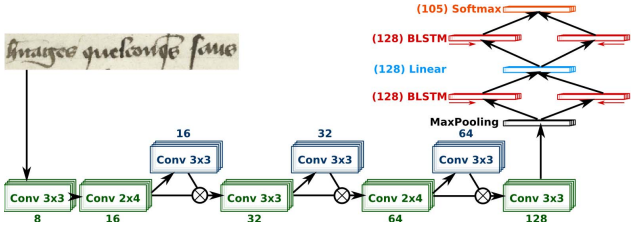


Fig. 1. Neural network architecture.

B. Character Lattices

For the "rich decoding" mentioned above, the character language model is represented as a weighted finite-state transducer. Beam search is then carried out by injecting in the transducer edges the scores predicted by the RNN, scaled with character priors [14]. The CLs are generated by the decoder implemented in the Kaldi toolkit [15], using beam search.

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 [8] (see also [5]), it provides a good approximation to the joint probability distribution $p(\mathbf{c}, \mathbf{x})$.

III. Probabilistic Framework for KWS

Let $\mathbf{R}$ be a random binary variable to denote whether an image region $\mathbf{x}$ is relevant for a query keyword v, formed by the concatenation of L characters $c_1, c_2 \ldots, c_L \stackrel{\text{def}}{=} \mathbf{c}_v$. Let $p(\mathbf{c}, \mathbf{x})$ be the joint probability distribution which, as discussed in Sec. II, is approximated by means of CLs obtained from the HTR decoding process.

As explained in [5], the probability that $\mathbf{x}$ be relevant for the query $\mathbf{c}_v$ is computed as:

$$P(\mathbf{R} \mid \mathbf{c}_v, \mathbf{x}) = \sum_{\mathbf{c} \in \Sigma^* \mathbf{c}_v \Sigma^*} P(\mathbf{c} \mid \mathbf{x}) = \sum_{\mathbf{c} \in \Sigma^* \mathbf{c}_v \Sigma^*} \frac{p(\mathbf{c}, \mathbf{x})}{p(\mathbf{x})} \quad (1)$$

where, for the sake of clarity, $\mathbf{R} = 1$ has been rewritten as $\mathbf{R}$.

From the two approaches proposed in [5] to perform the computations of Eq. 1, here we adopt the faster one, identified as "posteriorgram-like based". In this method, $P(\mathbf{R} \mid \mathbf{c}_v, \mathbf{x})$ is approximated by the length-normalized maximum of the *frame-level character sequence score*, $S(\mathbf{c}_v, \mathbf{x}, i)$:

$$P(\mathbf{R} \mid \mathbf{c}_v, \mathbf{x}) \approx \max_{1 \le i \le M} S(\mathbf{c}_v, \mathbf{x}, i)^{\frac{1}{L^\sigma}} \quad (2)$$

where i denotes a horizontal position (or *"frame"*) in $\mathbf{x}$, M is the length (number of frames) of $\mathbf{x}$ and σ is a length-normalization weight, tuned empirically.

The posteriorgram-like score, $S(\mathbf{c}_v, \mathbf{x}, i)$, can be efficiently computed by means of the *backward*, $\beta(q)$, and *forward*, $\alpha(q)$, accumulated scores over each CL state q (q_I denotes the initial state) [5]:

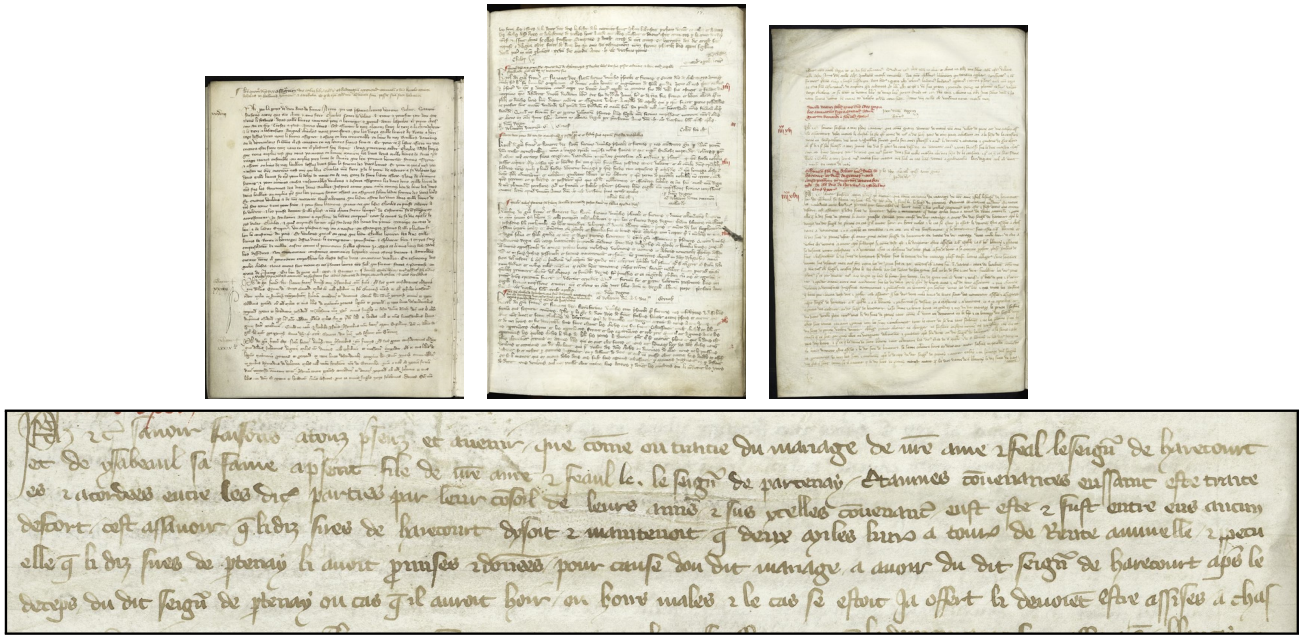
$$S(\mathbf{c}_v, \mathbf{x}, i) \stackrel{\text{def}}{=} \sum_{\substack{\mathbf{e}: \mathbf{c}_v = \Omega(\mathbf{e}), \\ \mathbf{e} = (q'_1, q_1), \ldots, (q'_L, q_L), \\ t(q'_1) < i \le t(q_L)}} \frac{\alpha(q'_1) \prod_{i=1}^{L} s(q'_i, q_i)\, \beta(q_L)}{\beta(q_I)} \quad (3)$$

In this equation, $\mathbf{e}$ is a CL *sub-path* of length L (sequence of CL edges: $e_1, e_2, \ldots, e_L$) such that $\mathbf{c}_v = \Omega(\mathbf{e})$, $\Omega(\mathbf{e})$ denotes the character sequence associated with a given sub-path $\mathbf{e}$, $t(q)$ is the horizontal position on $\mathbf{x}$ associated to the CL state q, and $s(e)$ is the likelihood score associated to the CL edge $e = (q', q)$. Please see [5] for more details of this KWS method.

IV. Dataset, Query Sets and Assessment Measures

A. Dataset Description

The complete corpus considered in the HIMANIS project is called "Chancery". It encompasses about 70 000 images reproducing the collection of medieval registers produced by the French royal chancery (Paris, Archives Nationales, JJ 35 - JJ 211), dating from 1302 to 1483, and containing ca. 68'000 charters given by the king of France. This large and iconic collection bears witness to the rationalization of late medieval administration and is a key source to our understanding of medieval Europe and the rise of centralized nation state on the



Philippes, etc. Savoir faisons à touz presenz et avenir que, comme ou traitié du mariage de nostre amé et feal le seigneur de Harecourt et de Ysabeaul, sa fame a present, file de nostre amé et feaul le seigneur de Partenay, certainnes convenances eussaint esté traite--es et acordées entre les dites parties par leur consoil de leurs amis, et sus ycelles convenances eust esté et fust entre eus aucun descort, c'est assavoir que li diz sires de Harecourt dysoit et maintenoit que deux miles livres à tournois de rente annuelle et perpetu--elle que li diz sires de Partenay li avoit promises et données pour cause dou dit mariage, à avoir du dit seigneur de Harecourt après le deceps du dit seigneur de Partenay, ou cas que il auroit hoir ou hoirs mâles, et le cas se estoit jà offert, li devoient estre assises à Chas-

Fig. 2. Examples of page images from the "Chancery" dataset. Extract from Paris, Archives Nationales, JJ 67, fol. 34v, n. 97 (1 May 1329). The ground truth, 'modernized' transcript of the text in the bottom image is also shown.

continent as a consequence of the long lasting wars between France and England. Nevertheless, the very size of this corpus prevented until now scholars to study it as exhaustively as it deserves. A user-friendly access to the contents of this key resource will help increasing the knowledge about medieval history and promote ongoing research in comparative studies on state management and administration.

The monumental text edition provided by Paul Guérin [16] contains a (relatively small) set of transcripts of more than 1 770 acts from this vast collection. These transcripts were converted in XML-TEI format by the Ecole Nationale des Chartes[3]. The ground-truth (GT) for the present work encompasses 436 images of these acts, corresponding to the texts edited in the vol. 1, 2, and 3, which cover the registers from JJ 35 to JJ 91, dating from 1302 to 1361. This subset represents merely 0.67% of the complete Chancery corpus, but is largely representative of its diversity and the challenges it represents. Examples of page images of this subset, as well as a trimmed act image with its modernized transcript, are shown in Fig. 2.

In a first stage, text lines of the images of this subset were semiautomatically detected and aligned with the GT transcripts. The manual corrections, carried out using Transkribus[4], helped uncovering a significant number of GT errors in the modernized text; nevertheless, some errors remain marginally, as well for the text and as for the alignment.

[3]http://corpus.enc.sorbonne.fr/actesroyauxdupoitou/.

[4]https://transkribus.eu.

Basic statistics of the dataset finally used for experimentation appear in Tab. I.

TABLE I. BASIC STATISTICS AND PARTITION OF THE CHANCERY SUBSET USED FOR EXPERIMENTATION. DC-FOLDED MEANS DIACRITICS- AND CASE-FOLDED

	Training	**Test**
Number of Acts	341	95
Number of Lines	6 061	1 733
Running Words	117 709	33 097
Lexicon Size	19 809	8 019
Lexicon Size (DC-folded)	15 677	6 579
Running OOV words	–	3 673 (11.1%)
Different OOV words	–	2 236 (39.4%)

It is important to note that, as the full Chancery corpus, this subset is fully bilingual (the acts can be written in Middle French or in Latin, or even both) and the GT modernized transcripts are largely different from the heavily abbreviated original text. In order to determine adequate keywords (i.e., expanded abbreviations) to explicitly assess how the proposed methods deal with this challenge, a smaller subset of 21 acts, encompassing 233 text lines was selected. The language used in each line (French or Latin) was identified and the corresponding diplomatic transcript was produced. This way, both the abbreviated and expanded (modernized) forms of all the words in both languages were provided for analysis and experimentation. According to the analysis, 71% of the text lines are written in French, and 22% of the running words in these lines are abbreviated. The remaining 29% of lines are written in Latin, with 59% of the running words abbreviated.

B. Query Set Selection

Several criteria can be adopted to define a set keywords to be used in KWS assessment experiments. In this work we adopt one of the criteria most commonly adopted (see e.g. [17]), where most of the words seen in the test partition are selected as spotting queries. Besides being a reasonable choice from an application point of view, this allows the *mean average precision* (mAP) measure (see IV-C) to be properly computed. Specifically, all the words in the test partition, excluding numbers and punctuation marks are used, making a total of 6 506 (diacritics- and case-folded) query words. A large proportion of these keywords are expanded forms of words which in the images appear abbreviated in several ways.

On the other hand, a query set exclusively composed of expanded forms of abbreviated words was used in the experiments explicitly devoted to abbreviations. For such a query set, we selected all the 244 unique (expanded) abbreviations which appear both in the 21 diplomatically transcribed acts (Sec. IV-A) and in the test set.

C. Evaluation Measures

KWS effectiveness was measured by means of standard *recall* vs. *interpolated precision* [18] curves, which are obtained by varying a threshold to decide whether a probabilistic score $P(\mathbf{R} \mid \mathbf{c}_v, \mathbf{x})$ (Eq. (2)) is high enough to assume that a word v appears in $\mathbf{x}$. From these curves, the *average precision* (AP) and the mean AP (mAP) [19], [20] are computed to provide simple, scalar performance measures. In addition, we provide the *maximum recall* for which a "useful" precision ($\geq 10\%$ – denoted as MxRc_{10}) is still achieved.

D. Experimental Setup

1) HTR and Lattice Generation: Due to the lack of diplomatized transcript, the neural network was trained to predict *modernized* transcription, even though the images contain many abbreviated words.

The neural network was trained with the RMSProp method on minibatches of 8 examples, using a base learning rate of 0.001, to minimize the CTC cost function. We stopped the optimization procedure when the error on the development set did not decrease for 20 epochs.

The obtained network was subsequently used with the Kaldi decoder to produce CLs for the lines of the test set. We performed one decoding pass for three character language models: 0-gram, 3-gram and 5-gram. They were all estimated with Kneser-Ney discounting and N-gram pruning with a threshold of $1e^{-7}$. We set the decoding beam to 20.

2) Indexing Process: Before computing the relevance probabilities (Eq. (2)), a preprocessing step was applied to the CLs. It basically consisted in a twofold task: all the characters associated to edges of the CLs' were case-folded and all the diacritics were removed. By computing the relevance probabilities using the resulting diacritics- and case-folded CLs, a diacritic- and case-insensitive probabilistic index is obtained, as required in the HIMANIS project.

Once all CLs have been processed in this way, the *forward-backward* scores were computed according to Eq. (3). Finally, relevance probabilities were computed from these scores according to Eq. (2). The parameters σ appearing in Eq. (2), was set up to 1, which in previous experiments proved to be an appropriate setting when working with RNN-based lattices.

V. Results and Discussion

A. HTR Results

After training, the test-set character error rate achieved by the raw RNN is 18.0%. 37.0% of the errors are substitutions, 11.2% are insertions and 51.8% are deletions. The high number of deletions can be explained by the relatively high proportion of abbreviated words (cf. Sec. IV-A). Since the amount of diplomatic transcripts is too small, the network is trained on the modernized version, where the abbreviations are expanded in the text, hence different from the image.

However, since the proportion of abbreviated words is high in the corpus, the network manages to learn – to some extent – to automatically expand those words. Fig. 3 illustrates how the network correctly expanded most abbreviations; namely, *predicte*, *Domini*, *nostri*, *dictum*, *percepit*, *quamdiu* and *nostro*. Using the RNN+5-gram system, *tempore* is correctly expanded too – and also *liberationis*, since *liberacionis* is another correct modernized expansion of this abbreviation. The only incorrectly expansion in this example happened in the first word (*casione*). However, in this an other similar cases, even if the 1-best prediction fails to be a correct expansion, the CL obtained by the decoder still assigns significant probabilities to other alternative hypotheses, generally including the correct one. Fig. 6, below, shows examples of this outstanding capability of the proposed approach to KWS.

B. Main KWS results

Experiments were carried out to evaluate the impact of the character language model on the performance of proposed KWS method. For the *full* query set of 6506 keywords, Table II reports AP, mAP and MxRc_{10} figures obtained for CLs produced using different N-gram models, with $N \in \{0, 3, 5\}$.

TABLE II. KWS performance: Average Precision (AP), mean AP (mAP) and maximum recall at 10% precision (MxRc_{10}) for lattices generated with different language models (LM).

Query Set	Lattice type	LM	AP	mAP	MxRc_{10}
Full (6506 keywords)	WLs	0-gram	0.637	0.462	0.795
	CLs	0-gram	0.618	0.524	0.709
	CLs	3-gram	0.692	0.613	0.800
	CLs	5-gram	**0.750**	**0.680**	**0.854**
	1-best	5-gram	0.581	0.445	0.698
Abbreviations-only (244 keywords)	CLs	5-gram	**0.838**	**0.736**	**0.908**
	1-best	5-gram	0.686	0.522	0.813

As expected, all these figures improve with the N-gram order, which is is in line with the results reported in [17], [5]. This shows once again the importance of leveraging linguistic context for KWS.

For comparison purposes only, the first row in Table II shows the KWS performance figures obtained for word lattices (WLs) by applying the approach described in [8]. Because of the 11.1% OOV rate (see Table I), these figures are poor with respect to those of 3-gram and 5-gram CLs, which do not suffer from the OOV problem. Table II also reports KWS

Fig. 3. Examples of RNN predictions for a line image containing abbreviations (extract from Paris, Archives Nationales, JJ 44, fol. 93v, n 150). From top to bottom: ground-truth (GT), input image, 1-best raw RNN prediction (0-gram), and 1-best prediction using a 5-gram character language model (+LM). The underlined segments correspond to abbreviated parts (characters not appearing in the image). Character insertion and substitution errors are marked with red color and the red vertical bar symbol (|) is used to indicate character deletion errors.

performance figures for "degenerate lattices" consisting of a single, linear path with the plain 1-best hypothesis of the HTR recognizer – this is equivalent to plain-text keyword searching in the single-best HTR transcripts.

Interpolated recall-precision (R-P) curves corresponding to each character N-gram order are plotted in Fig. 4, along with the R-P single point corresponding to 1-best recognition hypotheses with character 5-grams.

C. KWS Results of Searching for Abbreviated Keywords

Fig. 5 shows R-P curves and the corresponding AP and mAP values for searching for 241 expanded abbreviations using an index obtained from 5-gram CLs. The single R-P point for 1-best transcripts and the curves for French-only and Latin-only abbreviations are also shown.

The relatively better results achieved for abbreviations-only queries, with respect to the full query set, can be explained by the much larger average number of training examples available for abbreviations (110 examples per abbreviated word), with respect to the corresponding average number of training examples for normal words (6 examples per word).

These results clearly show that the proposed methods very successfully cope not only with multilingual (Latin/French) nature of text images and queries, but they are also able to very accurately spot severely abbreviated forms of queries posed in terms of expanded keywords. Examples of this remarkable capability can be seen in Fig 6.

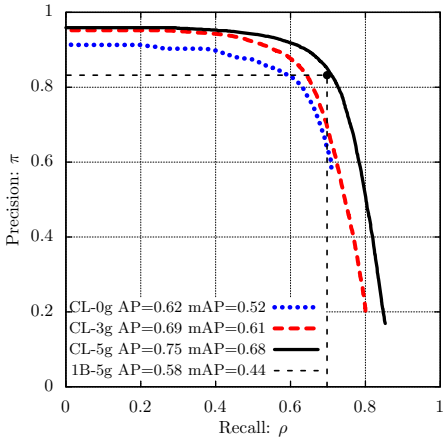


Fig. 4. *Recall-Precision* curves obtained for CLs using different character N-gram models (CL-Ng), for $N \in \{0, 3, 5\}$. A single R-P point (1B-5g) is also shown for the 1-best recognition hypotheses with character 5-grams.

VI. Concluding Remarks

Before undertaking the full, computationally intensive task of indexing the large historical manuscript collection aimed at in the HIMANIS project, preparatory experiments were carried out to assess the capabilities and expected performance of a relatively novel keyword spotting approach. In this approach word relevance image region probabilities are computed from character lattices produced by a holistic handwritten text recognition system based on recurrent neural networks and character N-gram language models.

Despite the extreme difficulties entailed by the large manuscript collection considered, results clearly show the feasibility of probabilistically indexing all the images of the collection, thereby making the iconic Chancery collection fully and accurately searchable by means of textual queries under the *precision-recall tradeoff model*.

In the time since the first version of this paper was written, most of the steps towards this goal have aleerady been done. This includes: a) carefully tune the size of the character lattices so that they still allow us not to significantly loose precision-recall performance, while reducing as much as possible the computing time needed to produce these lattices; and b) optimize the average number of word index entries produced per page image, in order to reduce the storage requirements as much as possible, while still maintaining essentially the original precision-recall performance.

Finally the whole collection has been actually indexed. The process required about 1 month of intensive multi-core computation and the resulting probabilistic index contains about 266 million entries and requires about 10 gigabytes

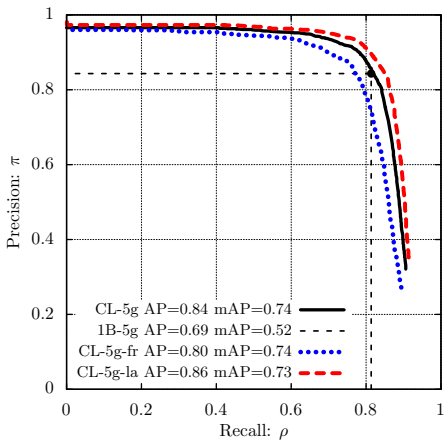


Fig. 5. *Recall-Precision* curve obtained with only abbreviated keywords for 5-gram CLs (CL-5g) and similar curves for French-only (CL-5g-fr) and Latin-only (CL-5g-la) abbreviations. A single R-P point (1B-5g) is also shown for the 1-best recognition hypotheses with character 5-grams.

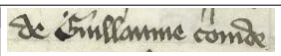
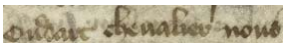
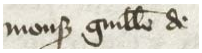
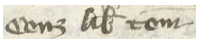
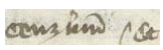
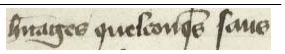
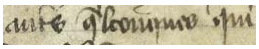
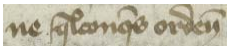
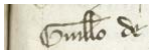
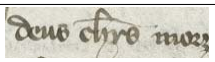
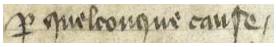
Keyword	**Guillaume**	**chevalier**	**livres**	**quelconques**
AP	0.79	0.89	0.79	0.91
t: d, h, m, f	0.45: 25,22,6,3	0.45: 39,34,3,5	0.20: 77,60,14,17	0.35: 11,10,1,1
π/ρ @ t	0.88 / 0.69	0.87 / 0.92	0.78 / 0.81	0.91 / 0.91
Full form				
Abbreviated	  	  	  	  
False Positives				

Fig. 6. Examples of modernized (expanded) keyword queries and corresponding spotting results. For each query: AP is the Average Precision, t is the confidence threshold used for the example images below, d is the number of detected spots, h is the number of correct hits, m is the number of missing spots, and f is the number of false positives, all at the threshold t. The precision and recall obtained from these values are π and ρ, respectively. Selected examples of correct spotted images, both in full form and abbreviated, and examples of false positives, are shown for each query.

of storage. During this process, about three million lattices were generated, then used to compute the probabilistic index entries, and finally discarded. All in all, this workflow involved handling about 250 gigabytes of data during the whole process time span of about two months. A beta version of the query and search system for the 67,282 page images of the full Chancery collection is available at prhlt-kws.prhlt.upv.es/himanis.

ACKNOWLEDGMENT

This work was partially supported by the Spanish MEC under FPU grant FPU13/06281, partially supported by the Generalitat Valenciana under the Prometeo/2009/014 project grant ALMAMATER, and through the EU projects: HIMANIS (JPICH programme, Spanish grant Ref. PCIN-2015-068) and READ (Horizon-2020 programme, grant Ref. 674943). Also, we thank Nvidia for the CPU Tesla K40c donation.

REFERENCES

[1] 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, special Issue on Awards from ICPR 2010.

[2] 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.

[3] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[4] 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.

[5] 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 International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Oct 2016, pp. 349–354.

[6] 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, no. 2, pp. 153–168, 2000.

[7] K. Terasawa and Y. Tanaka, "Slit style hog feature for document image word spotting," in *2009 10th International Conference on Document Analysis and Recognition*, July 2009, pp. 116–120.

[8] 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.

[9] M. Villegas, A. H. Toselli, V. Romero, and E. Vidal, "Exploiting existing modern transcripts for historical handwritten text recognition," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Oct 2016, pp. 66–71.

[10] A. Authors, "Gated Convolutional Recurrent Neural Networks for Multilingual Handwriting Recognition," in *International Conference on Document Analysis and Recognition*, 2017, p. (under review).

[11] 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.

[12] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *International Conference on Machine learning*, 2006, pp. 369–376.

[13] M. Federico, N. Bertoldi, and M. Cettolo, "Irstlm: an open source toolkit for handling large scale language models." in *INTERSPEECH*. ISCA, 2008, pp. 1618–1621.

[14] B. Moysset, T. Bluche, M. Knibbe, M. F. Benzeghiba, R. Messina, J. Louradour, and C. Kermorvant, "The a2ia multi-lingual text recognition system at the maurdor evaluation," in *International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 2014.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Workshop on Automatic Speech Recognition and Understanding (ASRU2011)*, 2011, pp. 1–4.

[16] P. Guérin and L. Celier, *Recueil des documents concernant le Poitou contenus dans les registres de la chancellerie de France*, ser. Archives historiques du Poitou. Poitiers: Oudin, 1881-1958.

[17] 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.

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

[19] M. Zhu, "Recall, Precision and Average Precision," Working Paper 2004-09 Department of Statistics & Actuarial Science - University of Waterloo, August 26 2004.

[20] S. Robertson, "A new interpretation of average precision," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR. New York, NY, USA: ACM, 2008, pp. 689–690. [Online]. Available: http://doi.acm.org/10.1145/1390334.1390453