Handwritten Text Recognition for the EDT Project. Part II: Textual Information Search in Untranscribed Manuscripts*

Enrique Vidal and Joan Andreu Sánchez

Abstract

Many massive handwritten text document collections are available in archives and libraries all over the world, but their textual contents remain practically inaccessible, buried behind thousands of terabytes of high-resolution images. If perfect or sufficiently accurate text image transcripts were available, image textual content could be strightforwardly indexed for plaintext textual access using conventional information retrieval systems. But fully automatic transcription results generally lack the level of accuracy needed for reliable text indexing and search purposes. And manual or even computer-assited transcription is entierely prohibitive to deal with the massive image collections which are typically considered for indexing. This paper explains the Probabilistic Indexing technology, which allows very accurate indexing and search to be directly implemented on the images themselves, without explicitly resorting to image transcripts. Results obtained using the proposed tecniques on several relevant historical data sets of the EDT project are presented, which clearly suppport the high interest of these technologies.

1 Introduction

In recent years, massive quantities of historical handwritten documents are being scanned into digital images which are then made available through web sites of libraries and archives all over the world. As a result of these efforts, many massive text *image* collections are available through Internet. The interest of these efforts not withstanding, unfortunately these document images are largely useless for their primary purpose; namely, exploiting the wealth of information conveyed by the text captured in the document images. Therefore, there is a fast growing interest in automated methods which allow the users to search for the relevant textual information contained in these images which is required for their needs.

In order to use classical text information retrieval approaches, a first step would be to convert the text images into digital text. Then, image textual content could be strightforwardly indexed for plaintext textual acces. However, as discussed in the first part of this publication (Sánchez and Vidal, 2021), OCR technology is completely useless for typical handwritten text images; and fully automatic, or even computer assited transcription results obtained using state-of-the art *handwritten text recognition* (HTR) techniques lack the level of accuracy needed for reliable text indexing and search purposes(Graves et al., 2009; Romero et al., 2012; Vinciarelli et al., 2004).

This situation raises the need of searching approaches specifically designed for large text *image* collections. In these approaches, indexing and search must be directly implemented on the images themselves, without explicitly resorting to image transcripts. On the other hand, rather than “exact” searching (as in plaintext), search has to be performed with a *confidence threshold*, somehow specified by the user as part of the query in order to meet the *precision-recall trade-off* which is considered most adequate in each query[1].

[*] The first part of this publication deals with model training and automatic transcription and will appear in (Sánchez and Vidal, 2021).

[1] Depending on the application, confidence thresholds can be specified more or less explicitly. For instance, in cases where the spotting results are provided in the form of ranked lists, the threshold is indirectly defined by the size of the list.

Clearly, such a confidence-based query model can not be properly implemented by just using conventional textual information retrieval methods on the noisy output of an automatic HTR system. Therefore, recognition techniques are needed which attach confidence scores to alternative word recognition hypotheses. Keyword spotting (KWS)[2] is a traditional way to address search problems within this framework. More precisely, KWS aims at determining locations on a text image or image collection which are likely to contain an instance of a queried word, without explicitly transcribing the image(s).

Traditional work on handwritten KWS assumed previous segmentation of the text images into words. However, word pre-segmentation is plainly impossible for millions of historical handwritten images of interest and, even in favorable cases, it is quite prone to errors (Manmatha and Rothfeder, 2005; Papavassiliou et al., 2010), which tend to hinder overall KWS performance significantly (Ball et al., 2006). To overcome this important drawback, recent works[3] assume the (word-unsegmented) *line image* as the lowest search level. This is a convenient setting because, in most cases of interest, text images can be fully automatically segmented into lines with fair accuracy (Bosch et al., 2012; Papavassiliou et al., 2010) and lines are sufficiently precise target image positions for most practical document image search and retrieval applications. Nevertheless, robust line segmentation can also be problematic in many cases and nowadays is considered perhaps the most severe bottleneck to achieve fully automatic processing of handwritten images for KWS and HTR alike. For this reason, our current work aims at indexing full pages, in an attempt to circumvent the need for any kind of image segmentation alltogether.

On the other hand, most of the techniques which have been proposed for KWS can be considered to belong to one of these two broad classes: *training-based* and *training-free*. Training-based KWS methods are generally based on statistical optical (and language) models and typically adopt the QbS paradigm. Conversely, most training-free techniques are based on direct (image) template matching and assume the QbE framework.

The approaches we follow are training-based and therefore need some amount (tens to hundreds) of manually transcribed images to train the required optical and language models. In addition they may benefit from the availability of collection-dependent lexica and/or other specific linguistic resources. Our target applications are those involving large handwritten collections, where the effort or cost to produce these resources will be more than rewarded by the benefits of accurately making the textual contents of these collections available for exploration and retrieval.

Traditional KWS technology has aimed at searching for a few (tens, hundreds, or maybe thousands of) *"key words"*, which the users should provide as those which are most interesting to search for information in the considered collection. In many cases, these keywords are assumed to be known beforehand.

Clearly, these assumptions go astray when very large collections of manuscripts are considered. In these cases, users can by no means pre-compile any reasonable list of "keywords", and the only adequate approach is to let the system itself "discover" the words which are likely to appear in the text images.

On the other hand, for very large image collections it becomes computationally unfeasible to build a system that not only process the images and discover likly writen words, but also searches for the arbitrary words the users happen to include in their queries. Therefore, the system's work has to be divided in two parts: First, in an *off-line, preprocessing* phase, likly words are hypothesized and adequately used to index the text images. Then, in an *on-line, search* phase, user's queries are analyzed and the indexed images are searched for the words included in the queries.

We call this approach and the corresponding technologies *Probabilisti Indexing* (PrIx). Results of this apporach and technologies for many interesting historic handwritten documents have been published in our recent publications[4]. Here, we will report new results on additional historical collections considered in the EDT project.

[2] See (Cao et al., 2009; Fischer et al., 2012; Frinken et al., 2012; Kamel, 2010; Manmatha et al., 1996; Puigcerver et al., 2016; Rath and Manmatha, 2007; Rodríguez-Serrano and Perronnin, 2009; Toselli and Vidal, 2013*a*; Toselli et al., 2016; Wshah et al., 2012).

[3] See (Fischer et al., 2012; Frinken et al., 2012; Kolcz et al., 2000; Terasawa and Tanaka, 2009; Toselli and Vidal, 2013*a*; Toselli et al., 2016; Wshah et al., 2012).

[4] See (Bluche et al., 2017; Lang et al., 2018; Puigcerver, 2018; Puigcerver et al., 2020; Toselli, Romero, Vidal and Sánchez, 2019; Vidal et al., 2020).

2 Probabilistic Indexing and Search

An overview of the ideas behind the indexing and search technology we are developping is presented in this section. As previously commented, this technology assumes the *precission-recall trade-off search model* which requires *word confidence scores* computed for adequate regions of the text images of interest.

Pixel-level word confidence scores: the "posteriorgram". A basic concept on which the proposed approach relies is the so called *pixel-level "posteriorgram"*. In a nutshell, it is a probability map computed for a given image X and a possible query word v. At each position (i, j) of X, the posteriorgram provides the posterior probability that the word v is written in some subimage of X which includes the pixel (i, j). Fig. 1 illustrates this concept.

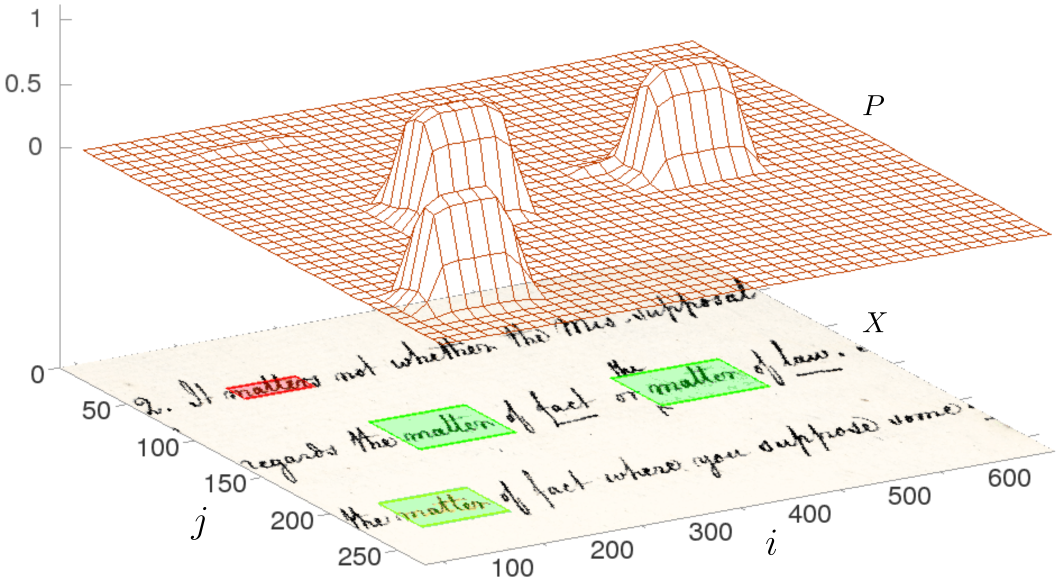

Figure 1: Pixel-level posteriorgram, P, for a text image X and word v ="**`matter`**". The most probable regions of X where v may appear according to P are marked in color boxes (red: low, green: high).

The value of P at each image position (i, j) can be easyly obtained by statistical *marginalization*. In simple words, the idea is to consider that v may have been written in any possible bounding box of the image X which includes the pixel (i, j). The marginalization process simply adds the word recognition probabilities for all these bounding boxes. This means that a posteriorgram could be simply obtained by repeated application of any word classification system capable of recognizing isolated (pre-segmented) words. Obviously, the better the classifier, the better the corresponding posteriorgram estimates.

Directly obtaining a full pixel-level posteriorgram in this way entails a formidable amount of computation. However, as it will be discussed later, it can be very efficiently computed by clever combinations of subsampling of the image positions (i, j) and adequate choices of the marginalization bounding boxes.

In our approaches we use full-fledged holistic HTR systems to compute the required isloated word probabilities. This allows us to take advantage of linguistic context to obtain very accurate word classification probabilities. In Fig. 1, a contextual word classifier based on a n-gram language model was used to compute P for the word "**`matter`**". This helped to achieve very low probabilities in a region of X around $(i = 100, j = 200)$, where a very similar (but different) word, "**`matters`**", is written. Clearly, according to the language model, the 2-grams "**`the matter`**" and "**`matter of`**" are very likely, thereby boosting the probability that the word "**`matter`**" is written in the correct image regions. Conversely, the 2-grams "**`It matter`**" and "**`matter not`**" are very unlikly, resulting in very low pixel probabilities in the image region where the different word "**`matters`**" is written (things would roughly be the other way around should the query word be "**`matters`**" instead).

Image region word confidence scores. Posteriorgrams could be directly used for KWS: Given a confidence threshold τ, a word v is just spotted in all image positions (i, j) where P is greater than τ. Varying the threshold, adequate *precision–recall* tradeoffs can be achieved. However, this naive idea is not feassible for large image collections, simply because indexing word confidences for every image pixel is obviously impossible. For PrIx, what we really need is the confidence that a word v is written within a pre-specified image region, such as a line, a column, or a full page, without explicitly taking into account exactly where the word is written in the region or how many instances of the word may appear in this region. In information retrieval terms, this is called *"relevance"*. That is, for each image region to be indexed, we need to obtain the probability that this region is *relevant* for the given query word.

Exactly computing relevance probabilities can become complex. Nevertheless, a very simple and intuitivelly apealing approach is to obtain the region relevance probability for a word v just as the maximum pixel-level probability for v over all the pixels of the region. For instance, if X in Fig. 1 is considered a region to be indexed, the probability that X is relevant for the the query "**`matter`**" is adequately approximated just the maximum of the four picks of the posteriorgram shown in this figure.

Choosing adequate minimal searchable image regions: line-level PrIx. In our work so far, line-shaped regions have been adopted as the lowest image element to be indexed. From the user point of view, lines are are sufficiently precise target image positions for most practical document image search and retrieval applications. On the technical ground, on the other hand, line-shaped image regions are particularly useful because they allow for efficient computation of posteriorgrams by adequately choosing the bounding boxes needed for the underlying marginalization process and by clever vertical subsampling of image positions.

Regarding the choice of *marginalization bounding* boxes needed to compute the posteriorgram, for a line-shaped image region, these boxes can be simply defined just by horizontal segmentation.

On the other hand, with line-shaped image regions *vertical subsampling*, in general, amounts to just guessing a proper line height and then runing a vertical-sliding window of this height with some overlap. Moreover, in many cases of interest, text lines are fairly regular and standard line segmentation techniques can yield accurate results. This allows to save computation cost and tends to increase accuracy.

Finally, and most improtantly, line-shaped text image regions typically contain most[5] of the relevant lingusitic context needed for precise computation of word classification probabilities using a language-model based recognizer, as discussed elsewhere.

Efficient computation of posteriorgrams and relevance probabilities. In our approaches, line-level posteriorgrams are very efficiently computed using *Word Graphs*, obtained as a byproduct of recognizing full line-region images with a full-fledged holistic HTR system based on *optical character models* and (N-gram) *Language Models*, as discussed in Part I of this publication (Sánchez and Vidal, 2021). When applied to a line-shaped image region, these systems can take full advantage of the lingusitic context which is present in the image to provide very accurate, word classification probabilities. On the other hand, a WG obtained in this way provides lots of alternative horizontal word-level segmentations. These segments directly define very adequate sets of bounding boxes, exactly as required by the marginalization process used to compute the posteriorgrams.

Line-region relevance probabilities are directly computed from the corresponding posteriorgrams, as explained above. Then, they can in turn be easily and consistently combined to obtain *page-level* relevance probabilities (. . . and so on for higher level indexing of *chapters*, *books*, etc.

Searching for words in probabilistically indexed images. Once the PrIx's of an image collection are available, textual search can be carried out using the pseudo-words and the corresponding geometric information spots contained om the PrIx spots. To this end, classical plaintext search techniques can in principle be applied. It is worth pointing out, however, that since PrIx's are generally very large (as compared with plain text), computing efficiency becomes a major concern.

[5] Most, but not all: linguistic context is obviously lost and the line boundaries. This problem is being considered towards upcoming developments of handwritten search and retrieval technologies.

On the other hand, PrIx easily allows complex queries that go far beyond the typical single keyword queries of traditional KWS. In particular, full support for standard multi-word boolean and word-sequence queries have been developed in (Toselli, Vidal, Puigcerver and Noya-García, 2019) and used in many PrIx search applications for large and huge collections of handwritten text documents[6].

Most of these applications also provide another classical set of handy free-text search tools; namely, *wildcard* and *approximate* (also called *"fuzzy"* or *"elastic"*) spelling. These tools are generally considered remarkably useful search assets in practice.

Probabilistic Indexing and Search Systems. Fig.2 show a typical textual information retrieval system based on PrIx. Its major components are:

- *"Probabilistic Indexing (PrIx)"*: Off-line pre-computation of PrIx's
- *"Ingest"*: Off-line creation of the actual database. Typically a simple and computationally cheap process
- *"Search engine and GUI"*: On-line user query analisys, find the requested information and present the retrived images.

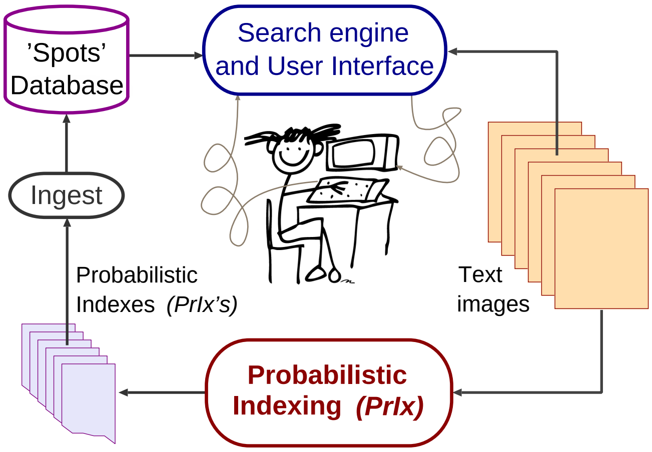
'Spots' Database
Search engine and User Interface
Ingest
Probabilistic Indexes (PrIx's)
Text images
Probabilistic Indexing (PrIx)

Figure 2: Probabilistic Text Image Indexing and Search System Diagram

PrIx is the most important component. As discussed above, it is based on *contextual word (or char string) recognition*, which require models *trained* from transcribed images (as in HTR). All this requires heavy (*off-line*) computing – but, as a result, it allows *extremely fast on-line query responses*, even for huge manuscript collections.

Another important component is the search engine and user interface, which should provide:

- *GUI*: Graphical / textual specification of queries and desired precision-recall tradeoff settings;
- *Query analysis*, which is trivial for single words, but becomes complex for multi-word queries, approximate spelling, etc.;
- *Search engine* to Access the database. Specialized software typically needed for probabilistically consitent support of multi-word queries, hierarchical search, and geometry-aware search;
- *Display retrieved images*: Prepare the images to be presented to the users as a result of their queries. The way they are presented is highly application dependent;
- Short response times are needed.

3 PrIx Search Evaluation Metrics

The standard *recall* and *interpolated precision* measures (Manning et al., 2008) are used to assess the effectiveness in all the search experiments.

[6] See `http://prhlt-carabela.prhlt.upv.es/PrIxDemos` for a list of PrIx live demonstrators.

For a given query and confidence threshold, *recall* is the ratio of relevant image regions (lines) correctly retrived by the system (often called "hits"), with respect to the total number of relevant regions existing in the image test set. *Precision*, on the other hand, is the ratio of hits with respect to number of (correctly or incorrectly) retrived regions.

By variing the confidence threshold, different related values of recall and precision can be obtained. These values can be ploted into the so-called *Recall-Precision* curve. Clearly, for a perfect system this curve would go stright from the point $(1, 0)$ vertically up to $(1, 1)$ and then horizontally left to $(0, 1)$. That is, such a system should exhibit a full precison (1) independently of the confidence threshold. This would in fact be the behaviour of a conventional plaintext retrieval system tested on a the perfect transcripts of the test set images. A reassonable search system should provide curves that go above the diagonal of the graph. the closer to the upper right corner (point $(1, 1)$), the better.

Results are also reported in terms of overall *average precision* (AP), which are obtained by computing the area under Recall-Precision curves and is a popular scalar assessment measure in Information Retrieval and KWS alike. Please refer to (Toselli et al., 2016) for details on these assesment measures.

4 Datasets

Many historical collections of handwritten text images have been considered in the past for testing the proposed indexing and search technologies. Most of the early work was carried out within the TRANSCRIPTORIM and READ projects mentioned in Sec. 1 of Part I of this paper (Sánchez and Vidal, 2021). Comprehensive accounts of these experiments have recently been reported in several publications[7].

Here we will present new experiments carried out through our collaboration with the EDT project. The features of these datasets are described in Part I of this paper. In each of these datasets, the set of query words used in the PrIx search experiments is the full set of words (vocabulary) of the test partition of each dataset. Specifically, the number of keywords used in each experiment are as follow: 657 for EDT-Hungary, 123 for EDT-Norway, 417 for EDT-Portugal, 322 for EDT-Spain and 273 for EDT-malta.

5 PrIx Search Results

In all the experiments the Optical and Language models used are trained from the training and validation partition of each dataset, as described in Sec. 2 and 6 of Part I of this paper (Sánchez and Vidal, 2021). PrIx search results were assessed using the recall-precision metrics outlined in Sec.3. A summary of the average precision (AP) for all the datasets is reported in Table 1 and the corresponding R-P curves are shown in Fig. 3.

Table 1: Average precision achieved in the EDT datasets.

EDT dataset	Hungary	Norway	Portugal	Spain	Malta
Average Precision (AP)	0.76	0.88	0.60	0.77	0.61

Results are reasonably good for all the datasets. As expected, the results are somewhat inferior for the more difficult collections. But even at these lower levels of performance, the systems can be used in practice to reliably find relevant information.

Overall these results are comparable to our previous results using the PrIx technology (see footnote)and very competitive as compared with results reported in the literature for classical KWS systems[8].

However, one may argue that these great laboratory results may not translate into a similarly good practical search experience. Consider for instance searching for information in the EDT-Spain collection. Typically, the

[7] See: (Bluche et al., 2017; Lang et al., 2018; Puigcerver, 2018; Puigcerver et al., 2020; Toselli, Romero, Vidal and Sánchez, 2019; Vidal et al., 2020)

[8] See (Fischer et al., 2012; Frinken et al., 2012; Rath and Manmatha, 2007; Rodríguez-Serrano and Perronnin, 2009; Toselli and Vidal, 2013*b*; Toselli et al., 2016; Wshah et al., 2012).

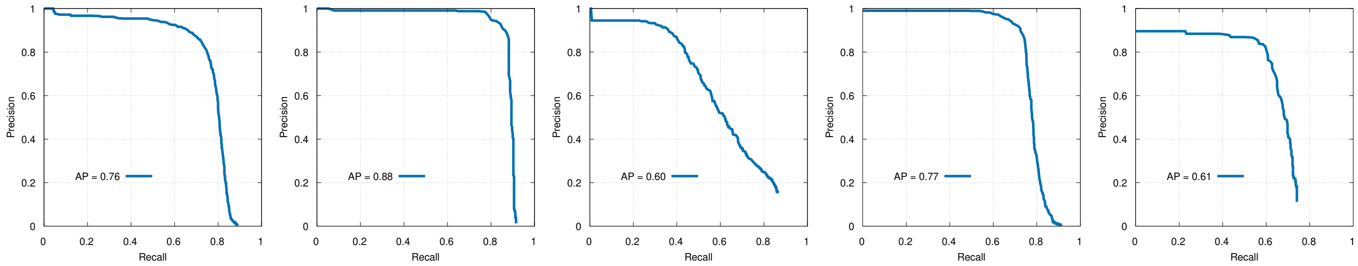

Figure 3: R-P curves for the EDT datasets. From letft to right: Hungary, Norway, Portugal, Spain and Malta.

user will try to find names of persons or cities, or maybe professions. In this scenario, an operational point such as RECALL$\approx$ 0.7 (and PRECISION$\approx$ 0.93, see Fig. 3) would fail to retrive an average of 30% of the spotscontaining the query word, while about 7% of the retrived lines would be false alarms.

Seveal factors, however, make the search experience much better than would be expected from these numbers. First, searching for information in handwritten text images can by no means be compared with conventional information retrieval, where no uncertainty exist about the words contained in the (electronic) documents searched for. In handwriten text images, the primary search baseline is just manual search; that is, visually scan each of the (maybe thousands or millions) page images, trying not to miss image regions where the query word does appear. Clearly, even an AP as low as 0.5 or even lower, may prove extreemly useful as compared with the manual baseline. Second, consider for instance the results of EDT-portugal dataset, with AP=0.60. These results are averaged over a query set of 452 diferent words. This set contained all the words seen in the test set, including frequent function words, and many other (short, more difficult to spot) words which are not typically query targets. For typical proper names, results are generally better (but experiments need to be done to objectively validate this assertion).

Finally, it worth to remind that, under the precision-recall tradeof search model, the user is not expected to be content with a fixed operational R-P point. Depending on the interest in finding only some, or most of the occurrences of a given query word, the users will try increasing or decreasing threshold values until they become satisfied with the results and/or understand they have meet the limitations of the system.

The real systems refered to in the next section can be used to get first-hand experience of the capabilities of these systems and the significancy of the results presented in this section.

6 Real PrIx and Search Systems for EDT Collections

Indexing and search engines similar to those used to obtain the results presented in Sec. 5, have also been used to support real search systems implemented according to the scheme of Fig. 2. These systems can be publicly accessed through Internet:

- EDT-Hungary: `http://edt.transkriptorium.com/hungary-search`
- EDT-Norway: `http://edt.transkriptorium.com/norway`
- EDT-Portugal: `http://edt.transkriptorium.com/por-tr`
- EDT-Spain: `http://edt.transkriptorium.com/esp`
- EDT-Malta: `http://edt.transkriptorium.com/malta`

Note that thses system are not exactly the same as those used in the experiments of the last section. The most important difference is that in these on-line systems, all the non-essential diacritics and special characters have been ignored and print/handwritten and semantic tags have been removed. In general, this can greatly sinplify the query experience and make it more effective. However, for collections such as EDT-Sapain, this actually hinders the system hability to honor complex, semantic-oriented queries and other advanced search capabilites, for which other systems have been set up, though they are not publicly available for the time being.

7 Conclusion and outlook

A formal probabilisitic framework has been introduced for indexing and searching large collections of handwritten documents. Empirical results with a variety of historic collections exhibiting differnet challenges and levels of complexity assess the usefulness of these methods in practice. Models trained for a given collection can provide quite useful performance on images from other similar collections, without need of (re-training). Several demonstrators have been implemented and made publicly available through the Internet for first-hand experience in real use.

In the coming future, work is planed to adress the folloing issues:

- So far line-regions are considered the most elementary elements to be indexed. This entails a requirement for automatic line detection and extraction. While fairly accurate automatic text line detection techniques exists, results lack robustness; that is, these techniques are not robust enough to reliably deal with the large variability in image quality and layout usually exhibited by historic handwritten documents. So, from time to time, a batch of page images appears in which line detection may fail dramatically. And, as a result, these pages become unindexed. Our current work aims at considering full page images as the lowest indexing level, in an attempt to completely circiunvent the line detection bottleneck (Barrere et al., 2019).
- All the techniques and experiments described in this paper assume that a user query is just a single word. Multiple word combined queries, and more specifically boolean and word sequence combinations (Toselli, Vidal, Puigcerver and Noya-García, 2019), as well as wild-card and approximate or flexible spelling are also supported in the real search systems. However, formal evaluation results for these complex queries still needs fundamental work do define adequate metrics and ecvaluation protocols.

8 Acknowledgments

References

Ball, G. R., Srihari, S. N., Srinivasan, H. et al. (2006), Segmentation-based and segmentation-free methods for spotting handwritten arabic words, *in* ‘Tenth Int. Workshop on Frontiers in Handwriting Recognition’.

Barrere, K., Toselli, A. H. and Vidal, E. (2019), Line segmentation free probabilistic keyword spotting and indexing, *in* ‘Iberian Conference on Pattern Recognition and Image Analysis’, Springer, pp. 201–217.

Bluche, T., Hamel, S., Kermorvant, C., Puigcerver, J., Stutzmann, D., Toselli, A. H. and Vidal, E. (2017), Preparatory KWS Experiments for Large-Scale Indexing of a Vast Medieval Manuscript Collection in the HIMANIS Project, *in* ‘Int. Conf. on Document Analysis and Recognition (ICDAR)’, Vol. 01, pp. 311–316.

Bosch, V., Toselli, A. H. and Vidal, E. (2012), Statistical text line analysis in handwritten documents, *in* ‘Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on’, IEEE, pp. 201–206.

Cao, H., Bhardwaj, A. and Govindaraju, V. (2009), ‘A probabilistic method for keyword retrieval in handwritten document images’, *Pattern Recognition* **42**(12), 3374–3382.

Fischer, A., Keller, A., Frinken, V. and Bunke, H. (2012), ‘Lexicon-free handwritten word spotting using character HMMs’, *Pattern Recognition Letters* **33**(7), 934 – 942. Special Issue on Awards from ICPR 2010.

Frinken, V., Fischer, A., Manmatha, R. and Bunke, H. (2012), ‘A Novel Word Spotting Method Based on Recurrent Neural Networks’, *IEEE Trans. on Pattern Analysis and Machine Intelligence* **34**(2), 211 –224.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H. and Schmidhuber, J. (2009), ‘A Novel Connectionist System for Unconstrained Handwriting Recognition’, *IEEE Transaction on Pattern Analysis and Machine Intelligence* **31**(5), 855–868.

Kamel, I. (2010), 'On indexing handwritten text', *Int. Journal of Multimedia and Ubiquitous Engineering* **5**(2).

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

Lang, E., Puigcerver, J., Toselli, A. H. and Vidal, E. (2018), Probabilistic indexing and search for information extraction on handwritten german parish records, *in* '2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)', pp. 44–49.

Manmatha, R., Han, C. and Riseman, E. (1996), Word Spotting: a New Approach to Indexing Handwriting, *in* 'Int. Conference on Computer Vision and Pattern Recognition (ICPR '96)', pp. 631–637.

Manmatha, R. and Rothfeder, J. L. (2005), 'A scale space approach for automatically segmenting words from historical handwritten documents', *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **27**(8), 1212–1225.

Manning, C. D., Raghavan, P. and Schtze, H. (2008), *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA.

Papavassiliou, V., Stafylakis, T., Katsouros, V. and Carayannis, G. (2010), 'Handwritten document image segmentation into text lines and words', *Pattern Recognition* **43**(1), 369–377.

Puigcerver, J. (2018), A Probabilistic Formulation of Keyword Spotting, PhD thesis, Univ. Politècnica de València.

Puigcerver, J., Toselli, A. H. and Vidal, E. (2016), 'Querying out-of-vocabulary words in lexicon-based keyword spotting', *Neural Computing and Applications* pp. 1–10.

Puigcerver, J., Toselli, A. H. and Vidal, E. (2020), Advances in handwritten keyword indexing and search technologies, *in* A. Fischer, M. Liwicki and R. Ingold, eds, 'Handwritten Historical Document Analysis, Recognition, And Retrieval-State Of The Art And Future Trends', Vol. 89, World Scientific, pp. 175–193.

Rath, T. and Manmatha, R. (2007), 'Word spotting for historical documents', *Int. Journal on Document Analysis and Recognition* **9**, 139–152.

Rodríguez-Serrano, J. A. and Perronnin, F. (2009), 'Handwritten word-spotting using hidden Markov models and universal vocabularies', *Pattern Recognition* **42**, 2106–2116.

Romero, V., Toselli, A. H. and Vidal, E. (2012), *Multimodal Interactive Handwritten Text Transcription*, Series in Machine Perception and Artificial Intelligence (MPAI), World Scientific Publishing.

Sánchez, J. A. and Vidal, E. (2021), Handwritten text recognition for the EDT project. Part I: Model training and automatic transcription, *in* M. A. Bermejo et al., ed., 'Proc. of the EDT Alicante workshop', To appear.

Terasawa, K. and Tanaka, Y. (2009), Slit style hog feature for document image word spotting, *in* 'ICDAR-09', pp. 116–120.

Toselli, A. H., Romero, V., Vidal, E. and Sánchez, J. A. (2019), Making two vast historical manuscript collections searchable and extracting meaningful textual features through large-scale probabilistic indexing, *in* '15th Int. Conf. on Document Analysis and Recognition (ICDAR)'.

Toselli, A. H. and Vidal, E. (2013*a*), Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents, *in* 'Proc. of the Int. Conf. on Document Analysis and Recognition (ICDAR'13)'.

Toselli, A. H. and Vidal, E. (2013*b*), Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents, *in* 'Proc. of the 12th Int. Conference on Document Analysis and Recognition (ICDAR '13)', IEEE Computer Society, Washington, DC, USA, pp. 501–505.

Toselli, A. H., Vidal, E., Puigcerver, J. and Noya-García, E. (2019), ‘Probabilistic multi-word spotting in handwritten text images’, *Pattern Analysis and Applications* **22**(1), 23–32.

Toselli, A. H., Vidal, E., Romero, V. and Frinken, V. (2016), ‘HMM word graph based keyword spotting in handwritten document images’, *Information Sciences* **370-371**, 497–518. Information Sciences 370-371 (2016) 497-518.

Vidal, E., Romero, V., Toselli, A. H., Sánchez, J. A., Bosch, V., Quirós, L., Benedí, J. M., Prieto, J. R., Pastor, M., Casacuberta, F., Alonso, C., García, C., Márquez, L. and Orcero, C. (2020), The carabela project and manuscript collection: Large-scale probabilistic indexing and content-based classification, *in* ‘17th Int. Conf. on Frontiers in Handwriting Recognition (ICFHR)’, pp. 85–90.

Vinciarelli, A., Bengio, S. and Bunke, H. (2004), ‘Off-line recognition of unconstrained handwritten texts using HMMs and statistical language models’, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(6), 709–720.

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