Schriften des Instituts für Dokumentologie und Editorik — Band 11

Kodikologie und Paläographie im digitalen Zeitalter 4

Codicology and Palaeography in the Digital Age 4

herausgegeben von | edited by

Hannah Busch, Franz Fischer, Patrick Sahle

unter Mitarbeit von | in collaboration with

Bernhard Assmann, Philipp Hegel, Celia Krause

2017
BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

Digitale Parallelfassung der gedruckten Publikation zur Archivierung im Kölner Universitäts-Publikations-Server (KUPS). Stand 4. September 2017.

SPONSORED BY THE

Federal Ministry
of Education
and Research

Diese Publikation wurde im Rahmen des Projektes eCodicology (Förderkennzeichen 01UG1350A-C) mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) gefördert.
Publication realised within the project eCodicology (funding code 01UG1350A-C) with financial resources of the German Federal Ministry of Research and Education (BMBF).

2017

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-7448-3877-1
Einbandgestaltung: Julia Sorouri, basierend auf Vorarbeiten von Johanna Puhl und Katharina Weber; Coverbild nach einer Vorlage von Swati Chandna.
Satz: LuaTEX und Bernhard Assmann

Advances in Handwritten Keyword Indexing and Search Technologies

Enrique Vidal

Abstract

Many extensive manuscript collections are available in archives and libraries all over the world, but their textual contents remain practically inaccessible, buried under thousands of terabytes worth of high-resolution images. If perfect or sufficiently accurate text-image transcripts were available, textual content could be indexed directly for plaintext access using conventional information retrieval systems. But the results of fully automated transcriptions generally lack the level of accuracy needed for reliable text indexing and search purposes. Additionally, manual or even computer-assited transcription is entierely unsustainable when dealing with the extensive image collections typically considered for indexing. This paper explains how accurate indexing and search commands can be implemented directly on the digital images themselves without the need to explicitly resort to image transcripts. Results obtained using the proposed techniques on several relevant historical data sets are presented, clearly supporting the considerable potential of these technologies.

Zusammenfassung

Auf der ganzen Welt halten Archive und Bibliotheken umfangreiche Sammlungen handschriftlicher Dokumente bereit. Doch bleiben deren Inhalte praktisch unzugänglich, verborgen unter tausenden von Terabytes hochaufgelöster Bilder. Gäbe es gute oder halbwegs verlässliche Text-Bild-Trankriptionen, ließen sich die jeweiligen Inhalte über herkömmliche Systeme zur Informationsrückgewinnung direkt indizieren und somit Zugänge zu entsprechenden Plaintext-Fassungen ermöglichen. Leider sind die Ergebnisse voll-automatisierter Transkriptionsverfahren zu ungenau, als dass sie sich für eine zuverlässige Textindizierung und Suche eigneten. Hinzu kommt, dass manuelle oder gar computergestützte Transkriptionsverfahren keine Nachhaltigkeit aufweisen, gerade wenn es sich um Bildsammlungen handelt, die aufgrund ihres großen Umfangs für eine Indizierung in Betracht gezogen werden. Dieser Artikel erläutert, wie verlässliche Indizierungen und Suchfunktionen unmittelbar auf den Bilddigitalisaten implementiert werden können, ohne dass dafür auf Bildtranskriptionen zurückgegriffen werden muss. Es werden Ergebnisse vorgestellt, die unter Anwendung der hier vorgestellten Technologie auf verschiedene historisch

Kodikologie und Paläographie im Digitalen Zeitalter 4 – Codicology and Palaeography in the Digital Age 4. Hrsg. Hannah Busch, Franz Fischer und Patrick Sahle, unter Mitarbeit von Bernhard Assmann, Philipp Hegel und Celia Krause. Schriften des Instituts für Dokumentologie und Editorik 11. Norderstedt: Books on Demand, 2017. 103–119.

bedeutsame Datensätze erzielt worden sind und deren erhebliches Potential klar unter Beweis stellen.

1 Introduction

Handwriting is, in a way, like speaking, but in contrast to spoken language the written word has the property that it does not vanish immediately as it is preserved in textual form. In the centuries since humanity discovered such a convenient way of persistent communication, large amounts of handwritten documents have been produced. In fact, it is argued that the accumulated amount of handwritten text so far is larger than the available amount of machine-written text today (copies excluded), including modern digital-born text. Notwithstanding the questionability of this conjecture, it is fairly probable that our current knowledge of the history of human societies, based on the infinitesimal amount of handwritten text that has painfully been transcribed, might be rather limited.

In recent years, large quantities of historical manuscripts have been digitised and made available through web sites of libraries and archives all over the world. As a result of these efforts, many massive *image* collections of textual documents are available online. Irrespective of these efforts and the interest in their products, unfortunately these digitisations are largely useless for their primary purpose: exploiting the wealth of information conveyed by the text captured in the images. Therefore, there is a fast growing interest in automated methods which allow users to search for relevant textual information contained in these images.

In order to use classical text information retrieval approaches, a first step would be to convert the text images into digital text. Then, image's textual content could directly be indexed for plaintext access. However, OCR technology is completely useless for typical handwritten text images - and the results of fully automated transcriptions obtained using state-of-the-art *handwritten text recognition* (HTR) techniques lack the level of accuracy needed for reliable text indexing and search purposes (Vinciarelli et al. 2004; Graves et al. 2009; Romero et al. 2012).

An alternative to fully automatic processing is to rely on *computer-assisted* transcription. This was successfully explored empirically by Toselli et al. (2017), Romero et al.(2012) and Alabau et al. (2014), following new, powerful concepts of pattern recognition-based human-machine interaction introduced by Vidal et al. (2007) and Toselli et al. (2011). Following the positive results of these laboratory studies, preliminary evaluation by real users was carried out by Toselli et al. (2016). In this case, a historical botany book of about one thousand pages was fully transcribed interactively in less than three months by a team composed of one paleographer and four paleography students. In the past four years, the tranScriptorium (Transcriptorium) project, has further explored the capabilities of these automatic and interactive HTR (IHTR) technologies to accelerate the conversion of raw text images into electronic text. These successful studies are now being continued within the recently started READ project.

Working conclusions from all studies mentioned above state:

a) To some extent, fully automatic transcripts of text images can be useful for plaintext indexing and search purposes. However, in many historical text image collections of interest, the typical level of transcription accuracy achieved severly hinders the search *recall*; i.e., the system's ability to ensure that all or most of the images which contain a given query text can actually be retrived is limited.
b) Similarly, the fully automatic transcription of most historical text images does not reach the level of accuracy needed for typical scholarly editions of the corresponding image collections.
c) In both cases, the required level of accuracy can obviously be obtained by means of additional user effort. If manual editing work is to be done, rather than just letting the users edit the noisy automatic transcripts, IHTR can be used to cost-effectively provide the desired transcription accuracy.
d) IHTR can significant reduce manual efforts regarding the edition of the automatic transcripts. But the overall effort demanded by IHTR is still substantial. Therefore, while IHTR is proving useful to produce scholarly editions of moderately sized historical collections, the required effort to handle extensive image collections targeted by indexing and search commands is entierely unsustainable.

This situation raises the need of search approaches specifically designed for large text *image* collections. In these approaches, on the one hand, indexing and search must be directly implemented in the images themselves, without explicitly resorting to image transcripts. On the other hand, rather than "exact" searching (as possible in plaintext), search queries have 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]

Clearly, such a confidence-based query model cannot be properly implemented just by using conventional textual information retrieval methods on the noisy output of an automatic HTR system. Therefore, recognition techniques are needed which attach confidence measures to alternative word recognition hypotheses. Keyword spotting (KWS)[2] is a traditional way to address search problems within this framework. More

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

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

precisely, KWS aims to determine locations on a text image collection which are likely to contain an instance of a queried word, without explicitly transcribing the images.

KWS is generally qualified as a Query-by-Example (QbE) or a Query-by-String (QbS), depending on whether the query word is specified by means of an example-image or just as a character string respectively. While the QbE scenario can be useful in some applications, it is clearly not adequate for our purposes of indexing and search in large image collections. Therefore, in this paper we adopt the QbS framework. Moreover, it has been shown by Vidal et al. (2015) that highly accurate QbS performance can be achieved easily by exclusively using QbS technology.

Traditional work on handwritten KWS assumed previous segmentation of the text images into word image regions. However, word pre-segmentation is impossible for millions of historical manuscript images of interest and, even in favorable cases, it is quite prone to errors (Manmatha and Rothfeder 2005; Papavassiliou et al. 2010) which generally result in poor overall KWS performance (Ball et al. 2006). To overcome this important drawback, recent works[3] assume the (non-segmented) *line image* as the lowest search level. This is a convenient setting because, in most cases, text images can be segmented fully automatically into lines with appropriate accuracy (Papavassiliou et al. 2010; Bosch et al. 2012) and lines are sufficiently precise as target image positions for most practical textual image search and retrieval applications. Nevertheless, a fixed line segmentation can also be problematic in many cases and is nowadays 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 proposed here are training-based and therefore need a certain amount (tens to hundreds) of manually transcribed images to train the required optical and language models. Additionally, they may benefit from the availability of collection-dependent lexicons and/or other specific linguistic resources. Our target applications are those involving large handwritten collections, where the effort or cost to produce these resources would pay off the benefits of making the textual contents of these collections readily available for exploration and retrieval.

Puigcerver et al. 2016; Toselli et al. 2016.

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

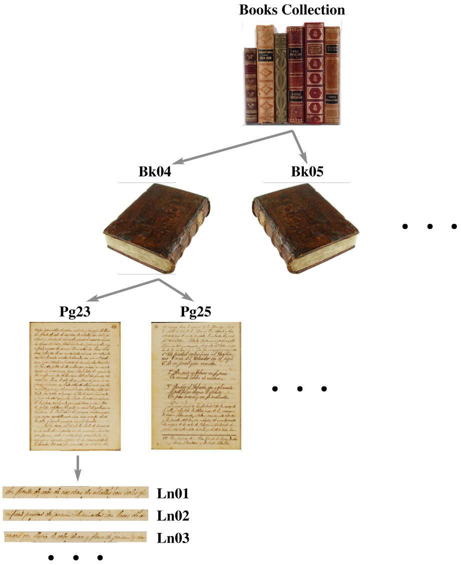


Figure 1: A hierarchical indexing and search model for handwritten text image collections. The top level in this illustration is a collection of books and the lowest level are line-shaped image regions. The specific levels of a hierarchy should be defined according to the characteristics of the document collection and search task considered.

2 Proposed indexing and search technology

An overview of the ideas behind the indexing and search technology we are developing is presented in this section. As previously stated, this technology assumes the *precision-recall trade-off search model* which requires *word confidence measures* computed for adequate regions of the text images of interest. Firstly, I will elaborate how these regions can be conveniently organized hierarchically and later I will explain how the required word confidence measures are computed.

A hierarchical indexing model

Indexing extensive document collections clearly calls for a hierarchical organization of indices. The lowest hierarchical level should consist of sufficiently small and meaningful *image regions*, such as text blocks (paragraphs) or lines. Figure 1 illustrates these concepts.

This kind of hierarchical organization of searchable text image regions entails important demands for the underlying precision-recall trade-off search model. Specifically, the word confidence measures must be defined properly, not only at the lowest level of the image region, but at every level of the hierarchy. In addition, con-

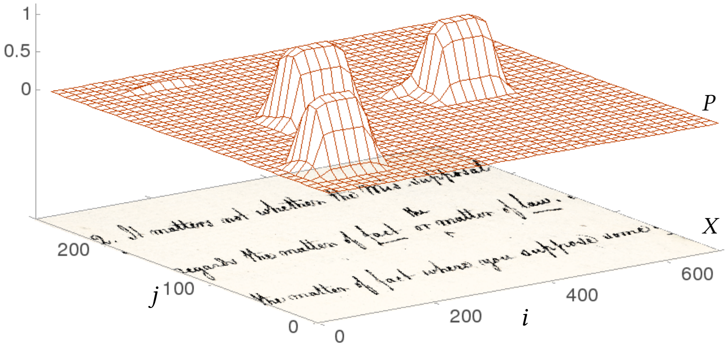


Figure 2: Pixel-level posteriorgram, P, for a text image X and word v = `"matter"`.

fidence measures must be properly normalized and homogeneous across hierarchy levels. Clearly, when a user is searching for a certain word his or her intuition about what a confidence level of e.g. 0.7 (70%) means should be the same whether he or she is searching for books in a book collection, for pages of a book, or for specific lines on one page of this book. This stands in direct opposition to the much less demanding confidence measure requirements entailed by conventional *flat indexing models*, which typically aim only to produce a ranked list of probable image regions retrived for each given query.

To fulfil the requirements discussed above, our techinques are being developed within a sound *statistical KWS framework* which supports the computation of confidence measures with the required properties, as explained below.

Pixel-level word confidence measures: the "posteriorgram"

The proposed approach relies on the basic concept of a *pixel-level "posteriorgram"*. In a nutshell, this 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). Figure 2 illustrates this concept.

The value of P at each image position (i, j) can be easily obtained by statistical *marginalization*. Simply put, 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 simply be obtained by repeated application of any word classification system capable of recognizing isolated (pre-segmented) words. It goes without saying, however, that 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 will be discussed later, it can be 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 fully fledged holistic HTR systems to compute the required isoloated word probabilities. This allows us to take advantage of the linguistic context to obtain accurate word classification probabilities. In figure 2, a contextual word classifier based on an n-gram language model was used to compute P for the word `"matter"`. This query led to comparatively low probabilities of X around $(i = 100, j = 200)$ in an region in which the similar (but different) word `"matters"` appears. According to the language model, the 2-grams `"the matter"` and `"matter of"` are highly predictable, thereby boosting the probability that the word `"matter"` exists in these exact image regions. Conversely, the 2-grams `"It matter"` and `"matter not"` are highly unlikely, resulting in low pixel probabilities in the image region where the different word `"matters"` appears (the results would roughly be reversed should the query word be `"matters"` instead).

Image region word confidence measures

Posteriorgrams can be used directly for KWS: given a confidence threshold τ, a word v is only spotted in image positions (i, j) where P is bigger than τ. Altering this threshold, adequate *precision–recall* trade-offs can be achieved. However, this approach is not feasible for large image collections as indexing word confidences for every image pixel would be impossible. For indexing purposes, 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 the exact location of a word in this specific region or the number of locations in which the word may appear. In information retrieval terminology, this is called *"relevance"*. For each image region to be indexed we need to obtain the probability that it is *relevant* for the given query word.

The process of exactly computing relevance probabilities can become a complex endeavor. Nevertheless, a comparatively simple and intuitively appealing approach is to compute the region relevance probability for a word v just as the maximum pixel-level probability for v over the whole region. For instance, if the whole X in figure 2 were considered a region to be indexed, the probability that X is relevant for the query `"matter"` is adequately approximated by the maximum of the four picks of the posteriorgram illustrated in this figure.

Choosing adequate minimal searchable image regions: line-level KWS

In our work so far, line-shaped regions have been adopted as the smallest and hierarchically lowest image elements to be indexed. From the user perspective, lines are target image positions sufficiently precise for most document image search and retrieval applications. From a technical perspective, on the other hand, line-shaped image regions are particularly useful as they enable efficient computation of posteriorgrams by adequately choosing the bounding boxes needed for the underlying marginalization process. Moreover, in many cases text lines are fairly regular and standard line segmentation techniques can be used to automatically determine line-shaped image regions with fair accuracy. Finally, and most importantly, line-shaped text image regions typically contain most[4] of the relevant lingusitic context needed for precise computation of word classification probabilities using a recognizer based on a language model, as discussed below.

Efficient computation of posteriorgrams and relevance probabilities
In our approach, line-level posteriorgrams are computed most efficiently using *Word Graphs* which are generated as a byproduct of recognizing full line region images with a fully-fledged holistic HTR system based on *optical character models* and (N-gram) *Language Models* (Toselli et al. 2016). When applied to a line-shaped image region, these systems can take full advantage of the lingusitic context to provide accurate word classification probabilities. On the other hand, a WG obtained in this manner provides a large number of alternative horizontal word-level segmentations. These segments directly define adequate sets of bounding boxes; just as those required by the marginalization process used to compute the posteriorgrams.

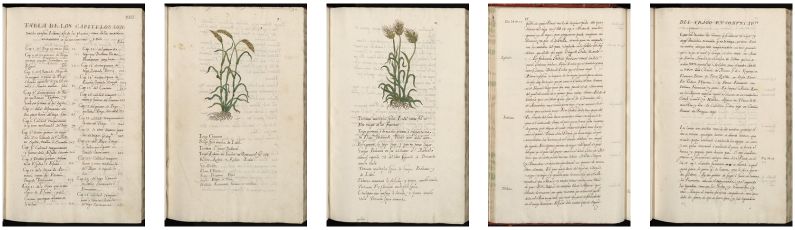
Line-region relevance probabilities are directly computed from the corresponding posteriorgrams, as explained above. They can in turn be combined easily and consistently to obtain *page-level* relevance probabilities (such as ... for *chapters*, *books*, etc., as needed for *hierarchical indexing*).

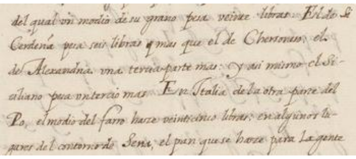
3 Laboratory results

Many collections of historical manuscript images have been considered for testing the proposed indexing and search technologies. Most of our research was pursued within the tranScriptorim project mentioned in section 1. The features of the data sets used in the experiments, the assessment measures adopted, and the results obtained, are presented in this section.

Data sets
A description summary and examples of the different data sets used in the experiments mentioned in this paper is given in figures 3–7. The first three data sets (Plantas, Bentham and Austen) correspond to collections which are comparatively modern (XVII-XIX century), entailing similar, relatively minor challenges in terms of writing style, homogeneity and language use. The last two data sets (Alcaraz and

[4] 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.





del qual vn modio de su grano pesa veinte libras: El de Si
Cerdeña pesa seis libras y mas que el de Cherroneso, el
del Alexandria vna tercia parte mas: y asi mismo el Si-
ciliano pesa vn tercio mas. En Italia de la otra parte del
Po, el modio del farro haze veinticinco libras; en algunos lu-
gares del contorno de Sena, el pan que se haze para la gente

Number of:	Total
Pages	881
Lines	19 764
Running words	196 858
Size of lexicon	20 931
Running characters	756 122
Size of character set	77

Figure 3: “PLANTAS”, XVII century Botanical Specimen Manuscript Collection of seven volumes written by a single writer in Old Spanish; page image examples and data set used for experiments on Vol. I.

WienSanktUlrich) correspond to more challenging early modern image collections exhibiting many of the difficulties entailed by medieval writing styles. The results of these laboratory experiments are presented in this section.

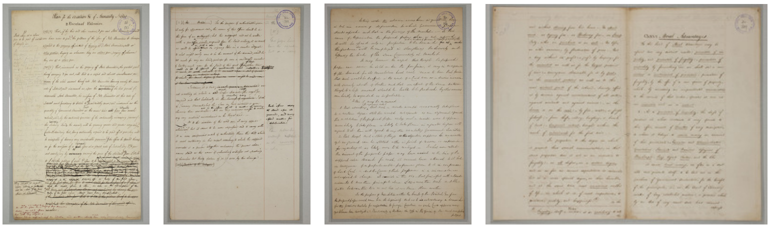
Further information regarding these data sets and the corresponding full collections can be found on the tranScriptorium web site (see section 1).

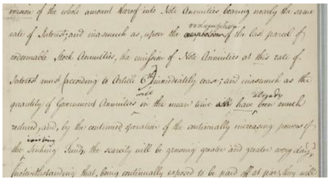
Evaluation measures

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

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 regard to the total number of relevant regions existing in the set of test images. *Precision*, on the other hand, is the ratio of hits with regard to the number of regions retrieved (both correctly or incorrectly).

By altering the confidence threshold, different related values of recall and precision can be obtained. These values can be plotted as a *Recall-Precision* curve. In a perfect system, this curve would rise vertically from point $(1, 0)$ to $(1, 1)$ before plateauing to $(0, 1)$. 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 perfect transcripts of the images of the test set. A reasonable KWS

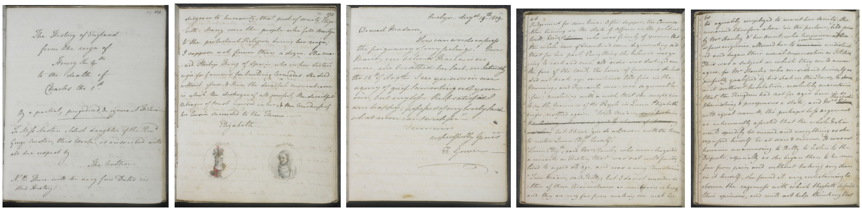


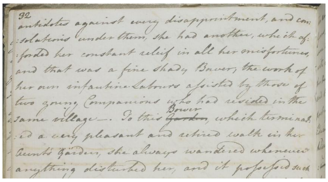


version of the whole amount thereof into Note Annuities bearing nearly the same
rate of Interest; and inasmuch as, upon the ~~completion~~ redemption of the last parcel of
redeemable Stock Annuities, the emission of Note Annuities at this rate of
Interest must (according to Article 6th) immediately cease; and inasmuch as the
quantity of Government Annuities will in the mean time ~~will~~ have already been much
reduced, and, by the continued operation of the continually increasing powers of
the ~~Sinking~~ operating Fund, the scarcity will be growing greater and greater every day;
notwithstanding that, being continually exposed to be paid off at par, they will

Number of:	Total
Pages	433
Lines	11 473
Running words	106 905
Size of lexicon	9 717
Running characters	550 674
Size of character set	86

Figure 4: “BENTHAM”, XVIII century collection of over 4, 000 volumes of drafts and notes, written in English by several writers; page image examples and data set of 433 selected page images used in the experiments.





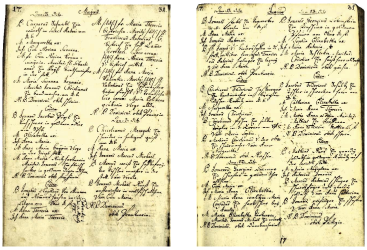
32
antidotes against every disappointment, and con:
:solations under them, she had another, which af:
:forded her constant relief in all her misfortunes,
and that was a fine shady Bower, the work of
her own infantine Labours aſsisted by those of
two young Companions who had resided in the
same village—. To this ~~Garden~~ Bower, which terminat:
:ed a very pleasant and retired walk in her
Aunt's Garden, she always wandered whenever
anything disturbed her, and it poſseſsed such

Number of:	Total
Pages	128
Lines	2 693
Running words	25 291
Size of lexicon	3 567
Running characters	118 881
Size of character set	81

Figure 5: “AUSTEN”, Jane Austen’s *Juvenilia*: XVIII century single hand manuscript in English; page image examples and “Volume The Third” data set used in the experiments.

- Close to 1000 page images; *miscellaneous hands, complex writing*
- About *30%* loosely *abbreviated words.*
- Experiments on 44 pages, cross-validation test
- *Lexicon & Query set*: approx. 3 400 keywords
- Training with *diplomatic transcripts*

Figure 6: "ALCARAZ", XVI century Spanish Inquisition trial against Pedro Ruiz de Alcaraz; example page images and details of the data set used in the experiments.



- Tens of thousands of two-column pages; *single hand*, but *mixed script compelex writing*
- Experiments on 52 pages, cross-validation test
- *Lexicon & Query set*: approx. 2 300 keywords
- Training with *diplomatic transcripts*

Figure 7: "WienSanktUlrich", XVI century German/Latin handwritten birth records (Wien). Example page images and details of the data set used in the experiments.

system should provide curves that rise beyond the graph's diagonal – the closer it gets to the upper right corner (point (1, 1), the better.

Results are also reported in terms of overall *average precision* (AP) and *mean AP* (mAP) obtained by calculating the area under the Recall-Precision curve. Both AP and mAP are popular scalar assessment measures for KWS.[5]

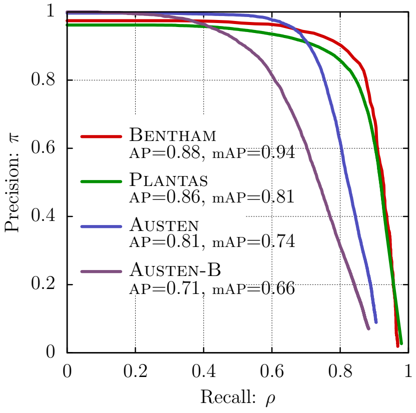
Results

Indexing and search results for the data sets described above are presented in figures 8 and 9. The results visualized in figure 8 correspond to the relatively modern (and less problematic) data sets (Plantas, Bentham and Austen). For the purpose of comparison, the results of figure 8 are also summarized (as gray curves) in 9, which mainly shows the results of the more challenging early modern data sets (Alcaraz and WienSanktUlrich)

In the case of the Austen data set, two experiments were carried out. In the first one, we adopted a conventional training-testing setting; i.e. KWS models were trained with annotated data of the same collection and performance was measured on an

[5] For details on these assessment measures see Toselli et al. 2016.

– *Recall-Precision* curves
– *Average Precision* (AP)
– *Mean Average Precision* (mAP)



Data sets training and test details

- **BENTHAM:** *miscellaneous hands. Training*: 400 pages from Bentham, 87 char. HMMs, 2-gram LM trained on Bentham texts; Lexicon 9 341 tokens.
  Test: 33 pages; query set: 6 962 keywords
- **PLANTAS (VOL-I):** *single hand. Training*: 224 pages from *Plantas*, 77 char. HMMs, 2-gram LM trained with the training set + book glossary transcripts. Lexicon 11 561 tokens.
  Test: 647 pages; query set: 9 945 keywords
- **AUSTEN:** *single hand. Training*: 50 Austen pages, 81 char. HMMs, 2-gram LM trained on Austen texts; Lexicon 20K tokens.
  Test: 78 pages; query set: 2 281 keywords
- **AUSTEN-B:** *single hand.* ***No training***; using Bentham character HMMs, lexicon and LM.
  Test & query set: Same as for **AUSTEN**

Figure 8: Results on XVII-XIX century manuscript image collections

independent test set of the same collection. In the second experiment (AUSTEN-B), we used models which had been trained with transcribed BENTHAM images (the same used for the experiment with the BENTHAM data set) to index the images of the test set of the AUSTEN collection. This experiment was aimed at exploring whether a handwritten image collection can be indexed fully automatically without previous training an that particlular collection by using KWS models previously trained with images of similar handwritting styles.[6]

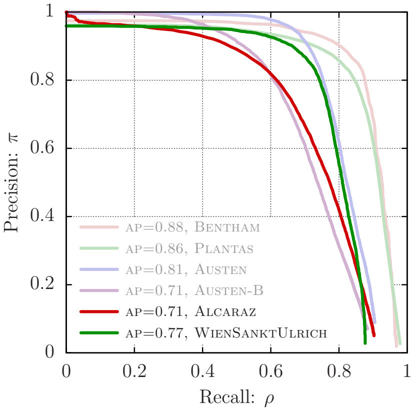
Good test results were achieved for all data sets. As expected, the results for the more difficult early modern data sets were less satisfying. However, even with this outcome the system can be used in practice to reliably find relevant information. The results for AUSTEN without training (i.e. using models trained for other, similar collections) were also somewhat inferior to those obtained with proper training with Austen data, but they still suffice to guarantee a successful use in practice.

Overall, the results presented above are competitive in comparison to results mentioned in the literature for classical KWS systems.[7] However, one may argue that these good laboratory results may not translate into a similarly satisfying prac-

[6] The writing style of AUSTEN was similar to the style of some of the writers of the BENTHAM collection (written in the same language and historical period).

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

– *Recall-Precision* curves
– *Average Precision* (AP)



Data sets training and test details

- **BENTHAM:** English, *miscellaneous hands. Training*: 400 pages; *Query set*: 6 962 keywrds
- **PLANTAS-I:** Spanish, *single hand. Training*: 224 p.; *Query set*: 9 945 keywords
- **AUSTEN:** English, *single hand. Training*: 50 pages: *Query set*: 2 281 keywords
- **AUSTEN-B:** English, *single hand.* ***No training*** (Bentham models). *Query set*: 2 281 keywords
- **ALCARAZ:** Spanish, *multi-hand. Training*: 44 pages, 70 char. HMMs, 2-gram LM trained on training transcripts; Lexicon 3 405 tokens. *Test*: Cross-val.; *Query set*: 3 400 keywords
- **WIENSANKTULRICH:**German/Latin, *one hand. Training*: 52 pages, 74 char. HMMs, 2-gram LM from training transcripts; Lexicon 2 303 tokens. *Test*: Cross-val.; *Query set*: 2 256 keywords

Figure 9: Results on early modern collections of manuscript images.

tical search experience. Considering, for instance, the search for information in the WIENSANKTULRICH collection, the user will try to find names of persons, cities, or possibly professions. In this scenario, an operational point such as `Recall`≈ 0.7 and `Precision`≈ 0.9 (see fig. 9) would fail to retrieve an average of 30% of the lines containing the query word, while about 10% of the retrieved lines would be false hits.

Several factors, however, contribute to a search experience much more positive than would be expected from these numbers. Firstly, searching for information in manuscript images can by no means be compared to conventional information retrieval where there is no uncertainty about the query words contained in the (electronic text) documents. In manuscript images the only approach generally available nowadays is a manual search; this entails visually scanning each of the (maybe thousands or millions) page images whilst trying not to miss image regions containing the query word. Here, even an Average Precision (AP) as low as 0.5 may prove extraordinarily useful in comparison to the basis of a manual search. Secondly, the results of figure 9 are averaged for a query set of 2 256 words. This set contains every token seen in training and in the test sets, including function words and many other words (shorter, more difficult to spot) which are not usually query targets. For proper names, results turn out generally better (but more experiments need to be conducted to objectively validate this assertion). Finally, given the precision-recall trade-off search model, the user is not expected to be content with a fixed operational point. Depending on the interest in finding only some, or most of, the occurrences of a given query word, the user will try increasing or decreasing threshold values until he or she is satisfied with the results and/or acknowledges to have met the limitations of the system.

The demonstration systems described in the next section can be used to gain first-hand experience of the capabilities of the systems in question and the significance of the results presented in this section.

4 Demonstration systems

The indexing and search engines used to obtain the results presented in section 3 are also used to support demontstration systems which can be publicly accessed online. Most of these demonstrators are available via the demonstrations section of the tranScriptoriun web site or via the following direct link:

http://transcriptorium.eu/demots/kws

It has to be be pointed out, however, that the demonstration for the PLANTAS collection does *not* include the first volume of the collection which was used in the laboratory experiments presented in the previous section.[8] In this case, the demonstrator is a working system proper, useful to find information in the about 1 000 pages of the *untranscribed* Vol. VII of the same PLANTAS collection. The optical and language models trained with pages of Vol. I and used to conduct the laboratory experiments were used to index the new, untranscribed Vol. VII fully automatically. Hence, this demonstrator can be seen as a typical and fitting example of the manifold possibilities provided by the technology presented in this paper.

5 Conclusion and outlook

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

[8] An older demonstration for Vol. I of PLANTAS is available at <http://cat.prhlt.upv.es/kws-demos>.

Future endeavors are planned to adress the following issues:

- So far, line-regions are considered the most fundamental elements to be indexed. This entails a requirement for automatic line detection and extraction. While there are fairly accurate automatic line detection techniques for textual data, results lack stability; these techniques are not stable enough to reliably tackle the significant variability in image quality and layout usually exhibited by historical manuscripts. Hence, at times a number of page images may appear in which line detection has failed drastically. As a result, these pages remain unindexed. Our current work aims at considering full page images as the lowest indexing level in an attempt to completely circumvent the line detection bottleneck.
- Techniques presented here require a predefined, possibly very large register of words to be indexed. Three approaches are currently being developped in order to overcome this limitation:
    - Probability *smoothing* techniques based on word similarities derived from character confusion probabilities
    - A *back-off* approach carrying out a computationally more extensive character-level search for queries involving non-indexed words
    - Do not longer insist in indexing *given* keywords; instead, find all the text elements which are likely to be “words” and just index all these “pseudo-words” blindly.
- All techniques and experiments described in this paper assume that a user query has the length of a single word. The development of techniques for multiple word and combined queries is currently in progress. Boolean and word sequence combinations in particular are already supported and formal evaluation results will be available in due course.

6 Acknowledgments

Thanks are due to Héctor Toselli, Joan Puigcerver and the HTR team at the PRHLT research center for their ideas and their work on the experiments illustrated in this paper. The author’s work was 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).

Bibliography

Alabau, Vincent, Carlos D. Martínez-Hinarejos, Verónica Romero, and Antonio L. Lagarda. "An iterative multimodal framework for the transcription of handwritten historical documents." *Pattern Recognition Letters* 35 (2014). 195–203.

Ball, Gregory R., Sagur N. Srihari, Harish Srinivasan et al. "Segmentation-based and segmentation-free methods for spotting handwritten arabic words." *Tenth International Workshop on Frontiers in Handwriting Recognition.* 2006.

Bosch, Vicente, Alejandro Héctor Toselli, and Enrique Vidal. "Statistical Text Line Analysis in Handwritten Documents." *Proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR'12)* 2012. 201–206.

Cao, Huaigu, Anurag Bhardwaj, and Venu Govindaraju. "A probabilistic method for keyword retrieval in handwritten document images." *Pattern Recognition* 42.12 (2009). 3374-3382. DOI: 10.1016/j.patcog.2009.02.003.

Fischer, Andreas, Andreas Keller, Volkmar Frinken, and Horst Bunke. "Lexicon-free handwritten word spotting using character HMMs." *Pattern Recognition Letters* 33.7 (2012). 934-942. DOI: 10.1016/j.patrec.2011.09.009.

Frinken, Volkmar, Andreas Fischer, R. Manmatha, and Horst Bunke. "A Novel Word Spotting Method Based on Recurrent Neural Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.2 (2012). 211-224. DOI: 10.1109/TPAMI.2011.113.

Graves, Alex, Marcus Liwicki, S. Fernández, Roman Bertolami et al. "A Novel Connectionist System for Unconstrained Handwriting Recognition." *IEEE Transaction on Pattern Analysis and Machine Intelligence* 31.5 (2009). 855–868.

Kamel, Ibrahim. "On Indexing Handwritten Text." *International Journal of Multimedia and Ubiquitous Engineering* 5.2 (2010).

Kolcz, Aleksander, Joshua Alspector, and Marijke F. Augusteijn. "A Line-Oriented Approach to Word Spotting in Handwritten Documents." *IEEE Transactions on Pattern Analysis & Applications* 3 (2000). 153–168. DOI: 10.1007/s100440070020.

Manmatha, Raghavan and Jamie L. Rothfeder. "A scale space approach for automatically segmenting words from historical handwritten documents." *Pattern Analysis and Machine Intelligence* 27.8 (2005). 1212–1225.

Manmatha, Raghavan, Chengfeng Han, and Edward M. Riseman. "Word Spotting: a New Approach to Indexing Handwriting." *1996 Conference on Computer Vision and Pattern Recognition (CVPR'96),* June 18-20, 1996. San Francisco (CA): IEEE, 1996. 631–637.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (eds.). *Introduction to Information Retrieval.* New York (NY): Cambridge University Press, 2008.

Papavassiliou, Vassilis, Themos Stafylakis, Vassilis Katsouros, and George Carayannis. "Handwritten document image segmentation into text lines and words." *Pattern Recognition* 43.1 (2010). 369–377.

Puigcerver, Joan, Alejandro H. Toselli, and Enrique Vidal. "Querying out-of-vocabulary words in lexicon-based keyword spotting." *Neural Computing and Applications* (2016). 1-10. DOI: 10.1007/s00521-016-2197-8.

Rath, Tony and Raghavan Manmatha. "Word spotting for historical documents." *International Journal on Document Analysis and Recognition* 9 (2007). 139-152.

READ: *Recognition and Enrichment of Archival Documents.* <http://read.transkribus.eu>.

Rodríguez-Serrano, José A. and Florent Perronnin. "Handwritten word-spotting using hidden Markov models and universal vocabularies." *Pattern Recognition.* 42 (2009). 2106-2116. DOI: 10.1016/j.patcog.2009.02.005.

Romero, Verónica, Alejandro Héctor Toselli, and Enrique Vidal (eds.). *Multimodal Interactive Handwritten Text Recognition.* New Jersey, London, Singapore et al.: World Scientific Publishing, 2012.

Terasawa, Kengo, and Yuzuru Tanaka. "Slit Style HOG Feature for Document Image Word Spotting." *Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'09) 2009.* 116–120.

Toselli, Alejandro Héctor et al. "Multimodal Interactive Transcription of Text Images." *Pattern Recognition* 43.5 (2010). 1814–1825.

Toselli, Alejandro Héctor, Enrique Vidal, and Francisco Casacuberta (eds.). *Multimodal Interactive Pattern Recognition and Applications.* London: Springer, 2011.

Toselli, Alejandro Héctor, Verónica Romero, Moisés Pastor-i-Gadea, and Enrique Vidal. "Transcribing a 17th century botanical manuscript: Longitudinal interactive transcription evaluation and ground truth production." *Digital Scholarship in the Humanities* 2017. DOI: 10.1093/llc/fqw064.

Toselli, Alejandro Héctor, and Enrique Vidal. "Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents." *Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR'13)* 2013. 501–505.

Toselli, Alejandro Héctor, Enrique Vidal, Verónica Romero, and Volkmar Frinken."HMM Word Graph Based Keyword Spotting in Handwritten Document Images." *Information Sciences* 370-371 (2016). 497-518. DOI: 10.1016/j.ins.2016.07.063.

tranScriptorium. 2013-2015. <http://transcriptorium.eu>.

Vidal, Enrique, Luis Rodríguez, Francisco Casacuberta, and Ismael Garcìa-Varea. "Interactive pattern recognition." *Proceedings of the International Workshop on Machine Learning for Multimodal Interaction.* London: Springer, 2007. 60–71.

Vidal, Enrique, Alejandro Héctor Toselli, and Joan Puigcerver. "High performance query-by-example keyword spotting using query-by-string techniques." *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR'15)* 2015. 741–745.

Vinciarelli, Alessandro, Samy Bengio, and Horst Bunke. "Off-line recognition of unconstrained handwritten texts using HMMs and statistical language models." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.6 (2004). 709–720.

Wshah, Safwan, Gaurav Kumar, and Venu Govindaraju. "Script Independent Word Spotting in Offline Handwritten Documents Based on Hidden Markov Models." *Proceedings of the International Conference on Frontiers in Handwriting Recognition. (ICFHR'12)* 2012. 14–19.