Making Large Collections of Handwritten Material Easily Accessible and Searchable

Anders Hast[1(✉)], Per Cullhed[2], Ekta Vats[1], and Matteo Abrate[3]

[1] Department of Information Technology, Uppsala University, Uppsala, Sweden
{anders.hast,ekta.vats}@it.uu.se
[2] University Library, Uppsala University, Uppsala, Sweden
per.cullhed@ub.uu.se
[3] Institute of Informatics and Telematics, CNR, Pisa, Italy
matteo.abrate@iit.cnr.it

Abstract. Libraries and cultural organisations contain a rich amount of digitised historical handwritten material in the form of scanned images. A vast majority of this material has not been transcribed yet, owing to technological challenges and lack of expertise. This renders the task of making these historical collections available for public access challenging, especially in performing a simple text search across the collection. Machine learning based methods for handwritten text recognition are gaining importance these days, which require huge amount of pre-transcribed texts for training the system. However, it is impractical to have access to several thousands of pre-transcribed documents due to adversities transcribers face. Therefore, this paper presents a training-free word spotting algorithm as an alternative for handwritten text transcription, where case studies on *Alvin* (Swedish repository) and *Clavius on the Web* are presented. The main focus of this work is on discussing prospects of making materials in the *Alvin* platform and *Clavius on the Web* easily searchable using a word spotting based handwritten text recognition system.

Keywords: Transcription · Handwritten text recognition · Word spotting · Alvin · Clavius on the Web

1 Introduction

Digital repositories offer a way to disseminate collections from archives, libraries and museums (ALM) on the web, where the general public can access the historical content with ease. Thus, it contributes to a world library scenario where human memory in the form of documents, books, photographs, letters, etc. is available and open to all.

The digital information is both similar and different from the original material. It is difficult to grasp a good idea of the materiality of the collections via the web, but on the other hand, completely new ways of working are available, based

P. Manghi et al. (Eds.): IRCDL 2019, CCIS 988, pp. 18–28, 2019.
https://doi.org/10.1007/978-3-030-11226-4_2

on the ability to search the digital material and process large amounts of data. These benefits widely outweigh the drawbacks and it should be a core task for everyone in the ALM sector to digitise and publish as much material as possible, as so much new knowledge for the benefits of humanity can be extracted from such collections.

This development is still in its early stages, and in this modern incunabula period of digital publications one could wish for, for example, that the handwritten material could be searched in a completely different way than is possible now. The digital repositories publish manuscripts and letters, but only as photographs of the original. Today, it is not possible to effectively translate the image of the handwritten text into a machine-readable text, in the way that can be done with printed books via Optical Character Recognition (OCR) techniques.

However, research in Handwritten Text Recognition (HTR) in advancing, and as researchers and companies work towards improvement of the HTR technology, the ALM sector must continue to publish the photographic images of the letters. This process also provides meta-data that is necessary to keep track of the material. In this regard, two popular projects *Alvin* and *Clavius on the Web* are discussed as follows.

1.1 Alvin

A Swedish repository, *Alvin* [1], consists of all groups of valuable digitised material that can be found within the ALM sector, which is unusual for digital repositories as they usually concentrate on a particular material category. *Alvin* is run by a consortium with the University libraries in Uppsala, Lund and Gothenburg at the centre, but several smaller archives and libraries also use *Alvin* for their digital publishing. In Alvin some entries contain only metadata, whereas the majority also contain digital images. At present, 163,000 entries are published and out of those 122,000 contain published images of the scanned books, manuscripts, drawings, maps etc. The metadata-only entries may contain unpublished images that are not visible due to, for example, copyright reasons. Among other things, *Alvin* contains tens of thousands of documents, manuscripts and letters published in digital form, but quite a few of them have been transcribed in their entirety. A war diary has been transcribed using dictation [2], and a number of documents from the Ravensbrück concentration camp have both been transcribed and translated, as in [3].

From a University library perspective, the major advantage of these documents at *Alvin* is that, due to Google indexing, their content is easily searchable on web. More people will find them and more people will use them.

There are several different repositories around the world that work with the same ambitions as *Alvin*. For example, Jeremy Bentham's archive at the University College of London has manually transcribed a large portion of Jeremy Bentham's diary pages [4]. Other archives with digitised material include, for example, the National Archives in Stockholm, which has digitised and published Court Protocols and Church Archives [5].

1.2 Clavius on the Web

Even when the main objective of a project or archive is not the publication and discoverability of the material per se, a manual transcription is often needed to enable further processing. It is the case of *Clavius on the Web* [6,7] and its sibling project *Totus Mundus*, aimed at enabling collaborative research for scholars, as well as teaching high school and undergraduate students. The Historical Archives of the Pontifical Gregorian University (APUG), upon which the initiatives are based, contain and preserve more than 5,000 manuscripts, written by Jesuits between 1551 and 1773. A portion of this wealth of material, alongside external resources such as ancient maps or computational tools, has been manually digitized and made available to the public [8,9]. Behind the scenes, scholars, technicians and students defined and developed together both the digital tools and artifacts to work with the original material, such as a domain-specific language for transcription [10] and computational lexica describing the text from a linguistic perspective [11,12]. Working within the development process and dealing with the transcription in a manual fashion was crucial for students. Nonetheless, the availability of an automatic technology for indexing and searching the digital images of the manuscript would have been of tremendous help in covering more portions of the archive and fostering understanding about its content. This is especially true for a technology that heavily relies on having a *human in the loop*, which would empower a student or researcher rather than trivialize the transcription work.

1.3 Other Projects

Several other projects work with manual transcription, often using the Omeka CMS and its Plug-in Scripto [13]. For instance, letters from the US Civil War at the Newberry Library in Chicago [14], or Moravian Lives [15], which is a collaboration between University of Göteborg and Bucknell University. However, the ALM world's way of solving the transcription issue has so far been to manually transcribe handwritten text in various ways. Manual transcription provides satisfactory results, but is rather too time consuming. Extremely few files have been made searchable through automatic transcription. The Monk system [16] is an example, but it can still be considered as a research project and not a generally useful way to automatically transcribe handwritten texts. The READ project [17] with the Transkribus software is a European project that used advanced HTR algorithms and has been used, for example, in transcribing Jeremy Bentham's archives. It has shown significant results whenever it is possible to train on a uniform material.

The amount of handwritten texts available in libraries and archives is enormous, and it will take a very long time before this material is transcribed and made searchable. However, large scale transcription can be accelerated potentially using holistic HTR techniques incorporated with expert user feedback, which is the main focus of this work.

This paper is organized as follows. Section 2 discusses related work on handwritten text transcription. Section 3 explains the proposed word spotting based HTR method in detail. Section 4 demonstrate preliminary results of using the word spotter on documents images from *Alvin* and *Clavius on the Web*. Section 5 concludes the paper.

2 Background

Transcription of historical handwritten documents is a tedious and time consuming task that requires skilled experts to manually transcribe lengthy texts. A technology driven alternative to manual transcription is to perform fully automatic transcription using HTR techniques that offer a rather cost-efficient solution. However, fully automatic transcription is often unreliable due to lack of validation of results [18], but it can be improved to deliver the desired level of accuracy by involving the user in the loop. This is often known as semi-automatic or semi-supervised transcription that is popularly used these days [18–22].

An interesting transcription technique, Computer Assisted Transcription of Text Images (CATTI), was proposed in [22] which is based on an interactive HTR technique for fast, accurate and low cost transcription. In general, it initiates an iterative interactive process between the CATTI system and the end-user for an input text line image to be transcribed. The transcription system hence generate significantly improved transcriptions by involving the end-user for providing corrective feedback.

A computer assisted handwritten text transcription approach was proposed in [20] that is based on image and language models from partially supervised data where the transcription algorithm employs Hidden Markov Models (HMMs) for text image modeling and n-gram models for language modeling. GIDOC (Gimp-based Interactive transcription of old text Documents) [23] system prototype has recently implemented this approach where a user navigates through transcription errors that are estimated using confidence measures generated using word graphs.

An active learning based line-by-line handwritten text transcription approach proposed in [21] continuously re-trains a transcription system to interact with an end-user to efficiently transcribe each line. In general, the performance of the above discussed transcription systems ([20–22]) highly depend on accurate detection and extraction of text lines in each document page. However, text line detection in old handwritten documents is a daunting task, that requires development of advanced line detection methods.

The Transcribe Bentham project [24] in collaboration with the READ project [17] is focused on generating handwritten text transcriptions using advanced HTR methods, where a significant amount of documents from Jeremy Bentham's collection have been transcribed using crowdsourcing that are used as training data for their machine learning based HTR algorithms. This innovative transcription technology is available through the Transkribus platform [25]. Similarly, the aforementioned Monk system [16] employs a HTR based word recognition system where volunteers help in labeling individual words through crowdsourcing and

hence help in generating the training data. However, such systems have limited applicability in a real world scenarios where pre-transcribed documents are not available to train a machine learning algorithm.

This work presents a segmentation-free word spotting algorithm, inspired from [26], for fast, efficient and reliable transcription of handwritten documents with little human effort, and is discussed in detail in the following section.

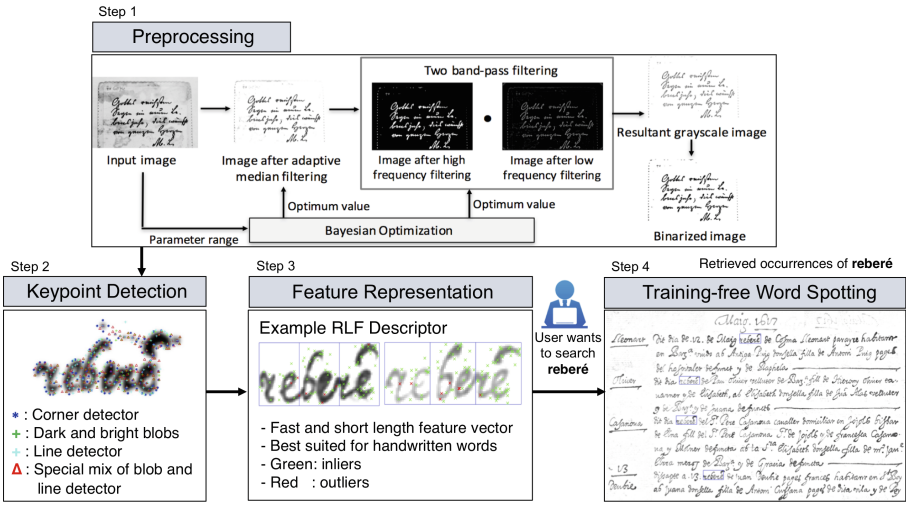


Fig. 1. General HTR pipeline involves efficient document preprocessing, keypoint detection, and feature representation before performing word spotting. If a user is interested in searching a word (say, *reberé*), all its occurrences on the document pages is retrieved as a result. (Color figure online)

3 Methodology

A training-free and segmentation-free word spotting approach towards document transcription is discussed where three important steps are formulated, as highlighted in Fig. 1. To begin with, in order to improve readability of poorly degraded manuscripts, an automatic document binarisation method was introduced in the previous work [27], which is based on two bandpass filtering approach for noise removal. With the help of Bayesian optimisation [28], best combination of hyperparameters are inferred by comparing the input image with its noise-free replica. This in turn generates an almost clean document image, free from noise such as due to bleed-through, contrast variations, wrinkles, faded ink etc. The reader is referred to *Step 1* of Fig. 1 for details on the preprocessing approach.

In general, the word spotter makes use of computer vision techniques for matching key points of the query word and a sliding window for fast recognition

of words. A combination of four different types of keypoint detectors was used (similar to [26]) to capture a variety of features that represent a handwritten text, and the keypoint detectors consisted of lines, corners and blobs (refer to *Step 2* of Fig. 1).

Recently, a Radial Line Fourier (RLF) descriptor with a short feature vector of 32 dimensions for feature representation of handwritten text has been proposed by the authors in [29]. Typically, existing feature descriptors such as SIFT, SURF, etc. inhibit invariant properties that amplify noise in degraded document images [30,31], and are found to be unsuitable for representing complex handwritten words with high levels of degradations. This inspired the authors to design the RLF descriptor to capture important properties of a handwritten text, and has been presented as *Step 3* in Fig. 1.

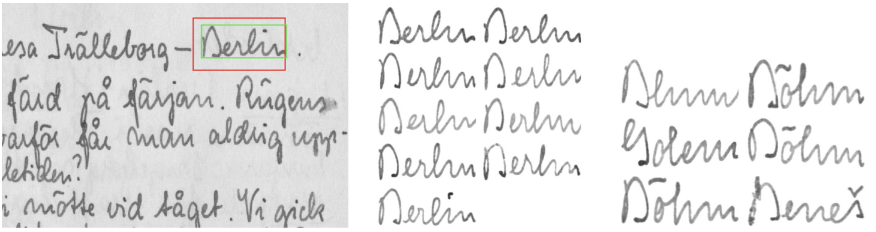
Using the RLF descriptor, the word spotting problem is reduced to a much faster word search problem, and referring to *Step 4* of Fig. 1, word spotting is performed as follows. The proposed system generates a document page query where a user is asked to mark a query word using a drag-and-drop feature that generates a rectangle bounding box (red in color). Furthermore, the algorithm automatically finds the best fitting rectangle (green in color) to perfectly encapsulate the word, and extracts the word as a result.

The words are typically partitioned into several parts, and a part-based preconditioner matching [29] is performed to avoid confusion between similar words and reduce false positives. This is because parts of the retrieved words may be similar to some part of the query word, or a word may share several characters with other words, and hence generate false positives [30].

The partitioning step is followed by a nearest neighbor search which is performed in an optimal sliding window within the subgroups of the detected keypoints. The extent of the matching points in a word is thus computed using a simple keypoint matching algorithm, that also captures words that are partially outside the sliding window. The resultant correspondences between the query word and the retrieved word needs further refinement due to the presence of potential outliers. To do so, a deterministic preconditioner [32] is used, and the matching algorithm efficiently captures complex variations in handwriting. An advantage of the proposed method is that it is completely learning-free, which means no prior knowledge about the text is required and word searching can be performed on a non-annotated dataset as well. *Alvin* and *Clavius on the Web* are an excellent use case to test the effectiveness of the proposed word spotter, and is demonstrated in the next section.

4 Experimental Framework

This section emphasize on the overall experimental framework of the proposed word spotter, and qualitatively evaluates the proposed method. The preliminary tests are performed on the images from the *Alvin* repository and the *Clavius on the Web*.



(a) User input (b) Found words (c) Rejected words

Fig. 2. Word spotting results for the query word *Berlin* for an example paper by a Swedish poet and novelist, *Karin Boye*, obtained from the *Alvin* portal [1]. Figure best viewed in colors. (Color figure online)

Figure 2 presents word spotting results on using a sample digitised image obtained from the *Alvin* portal of a travel diary by a Swedish poet and novelist, *Karin Boye*. In total, 15 pages by *Karin Boye* have been taken into account that belong to the period 1905–1943, written in Swedish language, and archived at

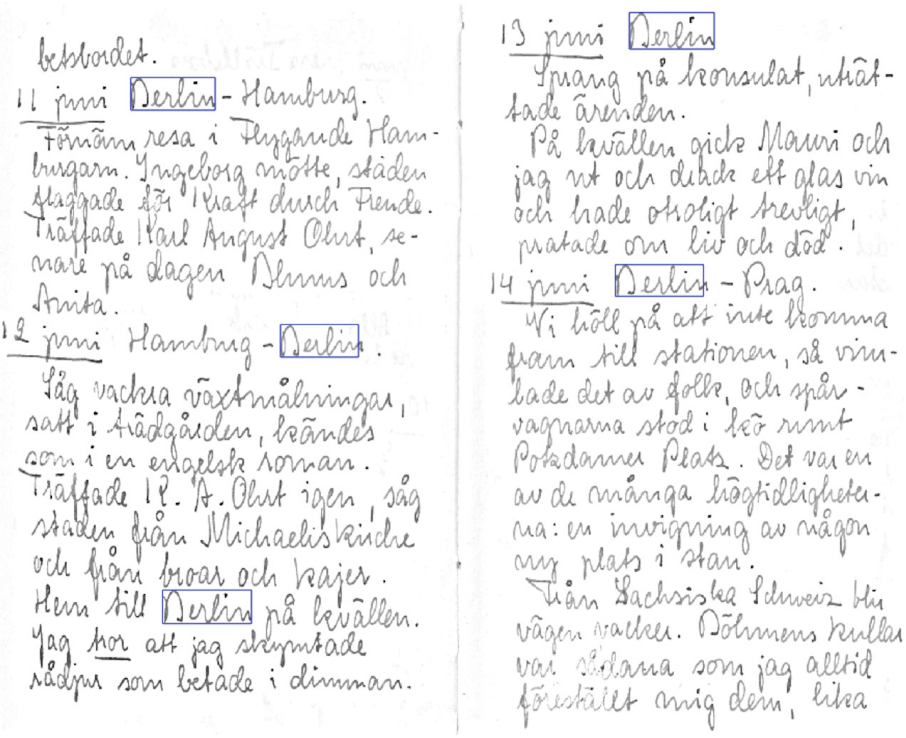
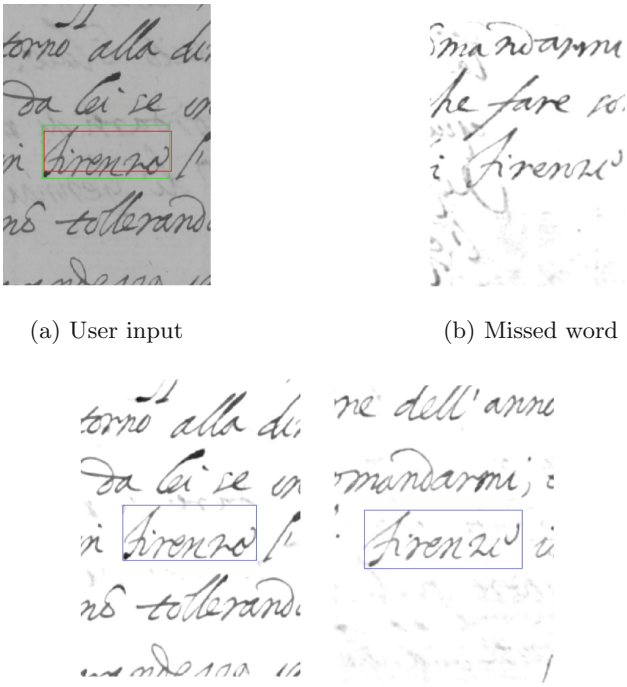


Fig. 3. Qualitative results obtained from the proposed training-free word spotter for the query word *Berlin*, for papers by *Karin Boye*, obtained from the *Alvin* portal [1]. Figure best viewed in colors. (Color figure online)



(a) User input (b) Missed word

(c) Found words

Fig. 4. Word spotting results for the query word *Firenze*, searching two letters written by *Galileo Galilei* to *Christopher Clavius*, obtained from the *Clavius on the web* portal [8]. The marked word (a) is found along with another occurrence of the same word (c), while one word is missed (b) due to the bleed-through problem. Figure best viewed in colors. (Color figure online)

Uppsala University Library. Referring to Fig. 2a, a query word *Berlin* has been marked by a user and denoted using red bounding box. The extent of the query word *Berlin* is automatically corrected by the algorithm, so that the word is perfectly encapsulated, and is represented using green bounding box. Figure 2b highlights several instances of the word *Berlin* found in the pages by *Karin Boye*, and it can be seen that all these founds words have been accurately identified by the proposed word spotter. Figure 2c presents some sample words that are very similar in characteristic with the word *Berlin*, and have been rejected by the word spotter (as intended). Furthermore, Fig. 3 presents some qualitative results obtained from the proposed training-free word spotter, with retrieved instances of the query word *Berlin* represented in blue bounding box.

The next set of experiments are performed on the digitised images of letters from the correspondence of *Christopher Clavius*, that are preserved by the Historical Archives of the Pontifical Gregorian University (Refer to the *Clavius on the Web* project [8]). For example, Fig. 4 highlights word spotting results for

letters written by *Galileo Galilei* in Florence 1588, to *Christopher Clavius* in Rome. In total, pages of two letters by *Galilei* have been taken into account. In Fig. 4a, a query word $firenze$ has been marked by a user (in red bounding box), automatically corrected to perfectly fit the word in green bounding box. Figure 4c presents the retrieved instances of the query word $firenze$ on 2 different letters, which have been accurately identified by the proposed word spotter. Interestingly, no word output has been rejected by the word spotter. However, there has been found an instance where the query word is missed by the word spotter due to high level of bleed-through in the document, and has been presented in Fig. 4c. This needs to be further investigated, and proposed word spotter can be further improved in future work. Usually, some amount of garbage words are found and we are currently investigating techniques to decrease the amount of such words that would only be irritating for the user. In the presented cases the algorithm was able to remove all such words, but unfortunately they are quite common, especially for shorter words. The words chosen here were names of cities as it belongs to the class of information usually interesting for historians and other users.

5 Conclusion

This paper presented a training-free word spotting method to facilitate fast, accurate and reliable transcription of historical handwritten documents available in digital repositories (such as *Alvin* and *Clavius on the Web* project). The word spotter efficiently finds multiple occurrences of a query word on-the-fly in a collection of historical document images. The preliminary experiments on sample images from *Alvin* and *Clavius on the Web* demonstrate the effectiveness of the proposed word spotter.

As future work, the aim is to develop a state-of-the-art comprehensive transcription tool to accelerate the time consuming transcription process using advanced HTR technology, incorporated with expert user feedback in the loop. The word spotting algorithm will in turn serve as a handwritten word search engine, similar to a Google for handwriting, where a user can search for a word query in historical archives in real-time.

Acknowledgment. This work was supported by the Swedish strategic research programme eSSENCE and the Riksbankens Jubileumsfond (Dnr NHS14-2068:1).

References

1. http://www.alvin-portal.org/ (2017)
2. http://urn.kb.se/resolve?urn=urn:nbn:se:alvin:portal:record-12537/
3. http://urn.kb.se/resolve?urn=urn:nbn:se:alvin:portal:record-100958
4. http://ucl.ac.uk/library/special-collections/a-z/ben-tham
5. https://sok.riksarkivet.se/digitala-forskarsalen
6. Abrate, M., et al.: Sharing cultural heritage: the clavius on the web project. In: LREC, pp. 627–634 (2014)

7. Pedretti, I., et al.: The clavius on the web project: digitization, annotation and visualization of early modern manuscripts. In: Proceedings of the Third AIUCD Annual Conference on Humanities and Their Methods in the Digital Ecosystem, p. 11. ACM (2014)
8. http://claviusontheweb.it
9. http://www.totusmundus.it
10. Valsecchi, F., Abrate, M., Bacciu, C., Piccini, S., Marchetti, A.: Text encoder and annotator: an all-in-one editor for transcribing and annotating manuscripts with RDF. In: Sack, H., Rizzo, G., Steinmetz, N., Mladenić, D., Auer, S., Lange, C. (eds.) ESWC 2016. LNCS, vol. 9989, pp. 399–407. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47602-5_52
11. Piccini, S., et al.: When traditional ontologies are not enough: modelling and visualizing dynamic ontologies in semantic-based access to texts. In: Digital Humanities 2016: Conference Abstracts, Jagiellonian University and Pedagogical University, Kraków (2016)
12. Piccini, S., Bellandi, A., Benotto, G.: Formalizing and querying a diachronic termino-ontological resource: the clavius case study. In: Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts, Proceedings of the Workshop, Krakow, Poland, 11 July 2016, pp. 38–41, no. 126. Linköping University Electronic Press (2016)
13. http://scripto.org/
14. https://publications.newberry.org/digital/mms-transcribe/index
15. http://moravian-lives.org/l
16. http://www.ai.rug.nl/~lambert/Monk-collections-english.html
17. http://read.transkribus.eu/
18. Romero, V., Bosch, V., Hernández, C., Vidal, E., Sánchez, J.A.: A historical document handwriting transcription end-to-end system. In: Alexandre, L.A., Salvador Sánchez, J., Rodrigues, J.M.F. (eds.) IbPRIA 2017. LNCS, vol. 10255, pp. 149–157. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58838-4_17
19. Terrades, O.R., Toselli, A.H., Serrano, N., Romero, V., Vidal, E., Juan, A.: Interactive layout analysis and transcription systems for historic handwritten documents. In: 10th ACM Symposium on Document Engineering, pp. 219–222 (2010)
20. Serrano, N., Pérez, D., Sanchis, A., Juan, A.: Adaptation from partially supervised handwritten text transcriptions. In: Proceedings of the 2009 International Conference on Multimodal Interfaces, ICMI-MLMI 2009, pp. 289–292. ACM, New York (2009)
21. Serrano, N., Giménez, A., Sanchis, A., Juan, A.: Active learning strategies for handwritten text transcription. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, ICMI-MLMI 2010, pp. 48:1–48:4. ACM, New York (2010)
22. Romero, V., Toselli, A.H., Vidal, E.: Multimodal Interactive Handwritten Text Transcription, vol. 80. World Scientific, Singapore (2012)
23. http://prhlt.iti.es/projects/handwritten/idoc/content.php?page=gidoc.php
24. Moyle, M., Tonra, J., Wallace, V.: Manuscript transcription by crowdsourcing: transcribe Bentham. Liber Q. **20**(3–4), 347–356 (2011)
25. http://transkribus.eu/Transkribus/
26. Hast, A., Fornés, A.: A segmentation-free handwritten word spotting approach by relaxed feature matching. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 150–155. IEEE (2016)

27. Vats, E., Hast, A., Singh, P.: Automatic document image binarization using Bayesian optimization. In: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing, pp. 89–94. ACM (2017)
28. Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., De Freitas, N.: Taking the human out of the loop: a review of Bayesian optimization. Proc. IEEE **104**(1), 148–175 (2016)
29. Hast, A., Vats, E.: Radial line Fourier descriptor for historical handwritten text representation. In: 26th International Conference on Computer Graphics, Visualization and Computer Vision (2018)
30. Zagoris, K., Pratikakis, I., Gatos, B.: Unsupervised word spotting in historical handwritten document images using document-oriented local features. IEEE Trans. Image Process. **26**(8), 4032–4041 (2017)
31. Leydier, Y., Ouji, A., LeBourgeois, F., Emptoz, H.: Towards an omnilingual word retrieval system for ancient manuscripts. Pattern Recognit. **42**(9), 2089–2105 (2009)
32. Hast, A., Marchetti, A.: An efficient preconditioner and a modified RANSAC for fast and robust feature matching. In: WSCG 2012 (2012)