Historical Document Processing: A Survey of Techniques, Tools, and Trends

James Philips and Nasseh Tabrizi
Department of Computer Science, East Carolina University, Greenville, North Carolina, U.S.A.

Keywords: Historical Document Processing, Archival Data, Handwriting Recognition, Optical Character Recognition, Digital Humanities.

Abstract: Historical Document Processing (HDP) is the process of digitizing written material from the past for future use by historians and other scholars. It incorporates algorithms and software tools from computer vision, document analysis and recognition, natural language processing, and machine learning to convert images of ancient manuscripts and early printed texts into a digital format usable in data mining and information retrieval systems. As libraries and other cultural heritage institutions have scanned their historical document archives, the need to transcribe the full text from these collections has become acute. Since HDP encompasses multiple sub-domains of computer science, knowledge relevant to its purpose is scattered across numerous journals and conference proceedings. This paper surveys the major phases of HDP, discussing standard algorithms, tools, and datasets and finally suggests directions for further research.

1 INTRODUCTION

Historical Document Processing (HDP) is the process of digitizing written and printed material from the past for future use by historians. Digitizing historical documents preserves them by ensuring a digital version will persist even if the original document is destroyed or damaged. Since many historical documents reside in libraries and archives, access to them is often hindered. Digitization of these historical documents thus expands scholars' access to archival collections as the images are published online and even allows them to engage these texts in new ways through digital interfaces (Chandna et al 2016; Tabrizi 2008). HDP incorporates algorithms and software tools from various subfields of computer science to convert images of ancient manuscripts and early printed texts into a digital format usable in data mining and information retrieval systems. Drawing on techniques and tools from computer vision, document analysis and recognition, natural language processing, and machine learning, HDP is a hybrid field. This paper surveys the major phases of HDP, discussing techniques, tools, and trends. After an explanation of the authors' research methodology, digitization challenges, techniques, standard algorithms, tools, and datasets are discussed, and the paper finally concludes with suggestions for further research.

2 METHODOLOGY

2.1 Research Rationale

This paper examines the evolution of the techniques, tools, and trends within the HDP field over the past twenty-two years (1998-2020). The authors believe this extended scope is warranted: No prior study was found that summarized the HDP workflow for both handwritten archival documents and printed texts. Prior studies have focused on one dimension of the problem, such as layout analysis, image binarization, or actual transcription. Very few discussed aspects of a full historical document processing workflow.

2.2 Article Selection Criteria

This research focuses on historical documents written in Latin, medieval and early modern European vernaculars, and English reflecting the current state of the HDP field: most of the work on historical archival documents has focused on western scripts and manuscripts. From the initial collection of 300+

Philips, J. and Tabrizi, N.
Historical Document Processing: A Survey of Techniques, Tools, and Trends.
DOI: 10.5220/0010177403410349
In *Proceedings of the 12th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2020) - Volume 1: KDIR*, pages 341-349
ISBN: 978-989-758-474-9

articles chosen, 50 were selected for this survey. This survey emphasizes the computer science dimension of HDP, especially machine learning methodologies, software tools, and research datasets. The authors envision other computer scientists, digital humanists, and software developers interested in HDP and cultural heritage as their primary audience.

3 TECHNIQUES AND TOOLS

3.1 Archival Document Types and Digitization Challenges

Historical documents broadly defined include any handwritten or mechanically produced document from the human past. Many have been preserved in the archives of museums and libraries, which have pursued extensive digitization efforts to preserve these invaluable cultural heritage artifacts. An enduring goal within the field of document image analysis has been achieving highly accurate tools for automatic layout analysis and transcription (Baechler and Ingold 2010).

A typical HDP workflow proceeds through several sequential phases.

Figure 1: The steps in a conventional HDP workflow for handwritten and printed documents.

After image acquisition, the document image is pre-processed and handwritten text recognition (HTR) or optical character recognition (OCR) is performed. This phase yields a transcription of the document's text. This transcription is the input to natural language processing and information retrieval tasks.

Prior to the 15th century, the majority of historical documents were texts produced by hand. After Gutenberg's printing press, published works were produced on the printing presses while private documents continued to be done by hand. This dichotomy in document types beginning in the Early Modern era led to diverse document types that must be dealt with differently during the HDP process.

The eclectic nature of all handwritten documents challenges automatic software tools. Medieval manuscripts are often more legible, and the inter-character segmentation of minuscule script are easier to train machine learning-based classifiers for than the continuous cursive of early modern handwritten texts. However, significant challenges in medieval documents are their complex layouts and intricate artwork (Simistira et al 2016). Continuous cursive script in Early Modern documents is challenging during the HTR phase, while medieval documents present greater challenges during layout analysis. Other challenges with historical documents include bleed-through from the opposite sides of the pages, illegible handwriting, and image resolution quality.

The earliest printed texts, known as incunabula, have posed the most difficulties for accurate, digital transcription of printed works (Rydberg-Cox 2009). Their fonts differ vastly from modern typefaces, and modern OCR software produces poor recognition results. The extensive use of textual ligatures also poses difficulties since they declined in use as printing standardized. After 1500 greater uniformity came to printed books, and by the early 19th century, the mass production of printed texts led to books that modern layout analysis and OCR tools could reliably and consistently digitize at scale, as seen in the digitation efforts of the Internet Archive and Google Books in partnership with libraries (Bamman and Smith 2012). This opens up possibilities for Information Retrieval in archival "Big Data."

3.2 Techniques

3.2.1 Pre-processing Phase

This pre-processing phase normally includes binarization/thresholding applied to the document image, adjustment for skew, layout analysis and text-line segmentation. Various studies have proposed various binarization methods including Bolan et al 2010, Messaoud et al 2012, and Roe and Mello 2013. Dewarping and skew reduction methods have been proposed in studies including Bukhari et al 2011 and performance analysis conducted in Rahnemoonfar and Plale 2013. Layout analysis is one of the most challenging aspects of HDP. Recent work has also examined the use of neural networks to restore degraded historical documents (Raha & Chanda 2019). Due to their complex page layouts, many studies have focused on layout analysis tools, algorithms, and benchmark datasets especially for medieval documents. Baechler and Ingold proposed a layout model for medieval documents. Using manuscript images from the E-codices project, they modeled a medieval manuscript page as several "layers": document text, marginal comments, degradation, and decoration. Overlapping polygonal

boxes are used to identify the constituent layers and are represented in software via XML.

Gatos et al 2014 developed a layout analysis and line segmentation software module designed to produce input to HTR tools. Their work was incorporated into the Transcriptorium project's Transkribus software.

Pintus, Rushmeier, and Yang likewise explore layout analysis and text-line extraction with an emphasis on medieval manuscripts. Pintus et al 2015 address the problem of initial calculation of text- line height. They segment the text regions coarsely and apply a SVM classifier to produce a refined text line identification. They note their method is not adversely affected by skewed texts and usually does not necessitate any alignment correction.

Yang et al (2017) extend their work on text-height estimation and layout analysis to an automated system that can work on a per-page basis rather than per manuscript. They propose three algorithms, one for text-line extraction, one for text block extraction, and one for identifying "special components." These use semi-supervised machine learning technique and focus on medieval manuscripts produced originally by professional scribes. Their results demonstrate that the desideratum of automatic algorithmically-layout analysis with high precision, recall, and accuracy is drawing nearer to reality.

3.2.2 Handwritten Text Recognition

Due to the inherent challenges of HTR for historical documents, some studies including (Rath and Manmatha 2006; Fischer et al. 2012) explored keyword spotting techniques as an alternative to producing a complete transcription. Early keyword spotting techniques approached it as an image similarity problem. Clusters of word images are created and compared for similarity using pairwise distance. Fischer et al explored several data-driven techniques for both keyword spotting and complete transcription (Fischer et al. 2009, 2012, 2014). One problem with word-based template matching is that the system can only recognize a word for which it has a reference image. Rare (out of vocabulary) words cannot be recognized. As a solution, the HisDoc project applied character-based recognition with Hidden Markov Model (HMM) to keyword spotting. For their keyword spotting analysis, they compared the character-based system with a baseline Dynamic Time Warp (DTW) system. Using Mean Average Precision as their evaluation metric, they found that the HMMs outperformed the DTW system on both localized and global thresholds for the George Washington and Parzival datasets (GW: 79.28/62.08 vs 54.08/43.95 and Parzival 88.15/85.53 vs 36.85/39.22). The HisDoc project also compared HMMs and neural network performance on the University of Bern's Historical Document Database (IAM-HistDB) to produce full transcriptions. They used a Bi-directional Long Short-term Memory (BLSTM) architecture that could mitigate the vanishing gradient problem of other neural network designs. Each of their nine geometric features used for training corresponds to an individual node in the input layer of the network. Output nodes in the network correspond to the individual characters in the character set. The probability of a word is computed based on the character-probabilities. According to (Fischer, Naji 2014), word error rates were significantly better for the neural network architecture than the HMM system on all three sets of historical document images: St. Gall 6.2% vs 10.6%, Parzival 6.7% vs 15.5%, and George Washington 18.1% vs 24.1%.

Neural networks continue to be the ascendant technique within the field for HTR. Granell et al. 2018 examined the use of convolutional recurrent neural networks for late medieval documents. The convolutional layers perform automatic feature extraction which precludes the need for handcrafted geometric or graph-based features such as those used by HisDoc. For deep neural network architectures to be competitive for time efficiency with other techniques, they require significant computational power. This is obtained through the use of a GPU rather than a CPU. Working with the Rodrigo dataset, they achieved their best results using a convolutional neural network supplemented with a 10-gram character language-model. Their word error rate was 14%.

3.2.3 Historical Optical Recognition

As with HTR, historical OCR can be accomplished with several techniques. However, neural network-based methods have become more prominent in the software libraries and literature recently. Since printed texts in western languages rarely use scripts with interconnected letters, segmentation-based approaches are feasible with OCR that are not practical for HTR. Nevertheless, historical OCR is drastically more difficult than modern OCR (Springmann and Lüdeling 2017). One challenge is the vast variability of early typography. Historical printings not laid out with modern, digital precision, and a plethora of early fonts were utilized across Europe (Christy et al 2017). A multitude of typeface

families exist, including Gothic script, Antiqua, and Fraktur. Although printing techniques standardized in the early 19th century, printed documents from 15th-19th centuries are too idiosyncratic for OCR machine learning classifiers trained using modern, digital fonts. Among the most difficult historical texts for OCR are incunabula due to their extensive use of ligatures, typographical abbreviations derived from medieval manuscripts that do not always have a corresponding equivalent in Unicode, and unpredictable word-hyphenation across lines (Rydberg- Cox 2009). The model training limitations of commercial software such as Abbey Fine Reader mean that researchers must resort to open source alternatives such as Tesseract or OCRopus (Springmann et al. 2014). Tesseract's classifier can be trained using either synthetic data (digital fonts that resemble historical ones) or with images of character glyphs cropped from actual historical text images. Tesseract and OCRopus both offer neural network classifiers Although high accuracies are achievable with neural networks, some of the same caveats apply from their use for HTR. These classifiers require substantial training data, with the corollary of extensive ground truth that must be created manually, and this classifier is computationally intensive for CPUs (Springmann et al 2014).

3.2.4 Software Tools and Datasets

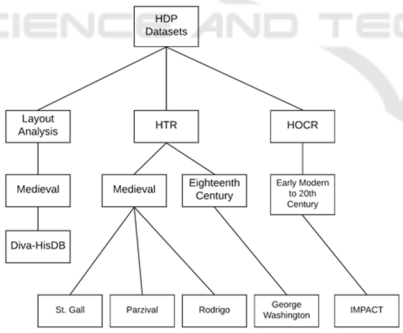

Figure 2: A taxonomy of HDP datasets based on use case and time-period.

Several software tools and datasets (Figure 2) exist for researchers and practitioners pursuing historical document processing. For historical OCR, these include Abbey FineReader, Tesseract, OCRopus, and AnyOCR tools and primarily the IMPACT dataset of early modern European printed texts. Few generic tools exist for historical HTR tasks, but researchers do have access to the IAM-HistDB and Rodrigo datasets. These variously contain images of full manuscript pages, individual words and characters, and corresponding ground truth for medieval Latin and early German and Spanish manuscripts. The IAM-HistDB also contains the Washington dataset for historical cursive handwriting recognition. In addition to software and datasets for the transcription phase of historical document processing, the Alethia tool and the IMPACT and Diva-HistDB datasets can be used for researching layout analysis and other pre-processing tasks. The rest of this section surveys the characteristics of the available datasets and discusses training, testing, and evaluation methodologies.

Few options exist for researchers seeking to work with medieval manuscript transcription. Two medieval datasets are included in the IAM-HistDB. The St. Gall dataset features images of a ninth century Latin manuscript written in Carolingian script by a single scribe. Fischer et al utilized the images and corresponding page transcriptions from J.P. Migne's *Patrologia Latina* previously published to create the dataset (Fischer et al. 2011). In addition to page images and transcription, the dataset includes extensive ground-truth: text-lines and individual word images have been binarized, normalized, and annotated with line-level transcription. Originally developed by the HisDoc project, the dataset has since been used in further research.

While Latin was the dominant ecclesiastical and scholarly language of Europe during the medieval period, some literature was produced in the vernacular languages. Two datasets exist for researchers investigating HTR in those vernacular texts, specifically the Old German and Old Spanish dialects. Included with the IAM-HistDB, the Parzival dataset contains manuscript pages of an Arthurian epic poem written in Old German from the 13th and 15th centuries. The 47 Parzival images are drawn from three different manuscripts produced by three scribes using Gothic minuscule script in multi-column layouts. Like the St. Gall set, the Parzival collection includes page images and transcription along with ground truth annotation. Text-lines and single word images have been binarized, normalized, and annotated with a full line-level transcription. Known as the Rodrigo corpus, the Old Spanish dataset is larger than either the St. Gall or Parzival datasets at 853 pages. Created for HTR and line extraction research, the researchers based at the Universitat Politecnica de Valencia used the digitized images of an Old Spanish historical chronicle, the "Historia de Espanana el archbispo Don Rodrigo" (Serrano et al 2011). The manuscript is from 1545, and thus can be traced to the the emergence of printing press technology. Although the creators of the dataset published results of running a hybrid

HMM-based image classifier with a language model, Granell et al have used the dataset with deep neural networks (Granell et al 2018).

The Washington dataset is the third dataset included in the IAM HistDB. Drawn from the George Washington papers at the US Library of Congress, its script is continuous cursive in the English language. First used in Rath and Manmatha, the HistDoc project supplemented the dataset with individual word and text-line images and corresponding ground truth transcriptions for each line and word (Fischer et al 2010). The Washington dataset is especially valuable for cursive HTR in historical documents.

The previously described IAM-HistDB datasets dealt exclusively with historical HTR. As a benchmark for evaluating pre-processing performance on medieval documents, the HistDoc project created the Diva-HistDB. This dataset contains 150-page images from three different manuscripts with accompanying ground truth for binarization, layout analysis, and line segmentation (Simistira et al 2016). Written in Carolingian script, two of the manuscripts are from the 11th century, and one from the 14th century written in Chancery script. All three manuscripts have a single column of text surrounded by extensive marginal annotation. Some pages have decorative initial characters. The layouts are highly complex. The ground truth concentrates on identifying spatial and color-based features. Like the IMPACT dataset, the ground truth is encoded in the PAGE XML format. The dataset is freely available on the HistDoc project website.

While most of the HTR and OCR datasets discussed in this section have focused on Latin languages or Latin script, a dataset has been created for HTR and OCR of historical polytonic (i.e. multiple accents) Greek texts. Introduced by Gatos et al, the dataset was developed for research on the word and character recognition as well as line and word segmentation (Gatos et al 2015). It features 399 pages of both handwritten and printed Greek text, mostly from the nineteenth and twentieth century.

3.2.5 Methodologies for Evaluation

Several metrics are used to evaluate the performance of a historical document processing system. For handwritten text recognition systems that use image similarity, precision and recall are two important performance measures. Precision ascertains how many of all the relevant results in the dataset were actually retrieved. For machine learning systems, transcription performance is evaluated using the character error rate, word error rate, or sometimes both if a language model is utilized to enhance the recognition results. Layout analysis performance is assessed using the line error rate and segmentation error rate (Bosch et al 2014).

3.2.6 Software Systems

Cultural heritage practitioners seeking production-ready tools for their own historical document preservation projects have two software systems available that provide a full suite of tools for pre-processing, machine learning training, and transcription. These two tools are DIVA-Services (Würsch et al 2017) and the Transkribus platform from the EU-sponsored READ project (Kahle et al 2017).

DIVA-Services and Transkribus offer similar feature sets to the cultural heritage community. However, they should not be seen as direct competitors. As a cross-platform software service, Transkribus is likely the better solution for archivists seeking an integrated HDP toolchain that requires minimal or no custom software to be developed. Since it offers multiple tools for each step in the HDP process and supports standard formats such as PAGE, it is ideally suited for archivists who need a reliable service for a historical document transcription project that allows support for machine learning training on new datasets. Due to the platform's hybrid open source-closed source nature and lack of tool modularity (users cannot substitute their own libraries directly for a Transkribus one), users who need more flexibility and alignment with open source values may find DIVA-Services more suited to their needs. Since DIVA-Services provides separate API calls for each discrete step in the HDP workflow, this service is more suitable for computer science researchers and archivists who need to integrate existing methods alongside custom software. DIVA-SERVICES and Transkribus thus offer complementary approaches that meet the different use cases of members of the cultural heritage community.

4 RECENT TRENDS

Within the past decade, several research projects have advanced the field of historical document processing through the creation of datasets, the exploration of improved techniques, and the application of existing tools to digital archival document preservation efforts. The HisDoc family of projects have made significant contributions to algorithms, tools, and datasets for medieval manuscripts. The inaugural

HisDoc project lasted from 2009 to 2013 and concurrently studied three phases of HDP: layout analysis, HTR, and document indexing and information retrieval (Fischer Nijay et al 2014). While much of their research focused on medieval documents and scripts, their goal was to create "generic methods for historical manuscript processing that can principally be applied to any script and language (83)."

HisDoc 2.0 was conceived as a direct extension of the original HisDoc project. Concentrated at the University of Fribourg, the focus of this project was advancing digital paleography for archival documents (Garz et al 2015). The HisDoc 2.0 researchers recognized that historical manuscripts are complex creations and require multi-faceted solutions from computer science. Written by multiple scribes and due to inconsistent layouts, many documents do not conform to the ideal characteristics explored during the first HisDoc project. With HisDoc 2.0, the researchers investigated combining text localization, script discrimination, and scribal recognition into a unified system that could be utilized on historical documents of varying genres and time periods. The HisDoc 2.0 project made several contributions to the field. One was DivaServices, a web service offering historical document processing algorithms with a RESTful (representational state transfer) API to circumvent the problem many developers and practitioners face with the installation of complicated software tools, libraries, and dependencies (Würsch et al 2016). Another contribution was the DivaDesk digital workspace, GUI-based software that makes computer science algorithms for ground truth creation, layout analysis, and other common tasks accessible for humanities scholars (Eichenberger et al 2014). The project explored ground truth creation, text region and layout analysis with neural networks, and aspects of scribal identification. Finally, the project produced and released the Diva-HisDB dataset.

The IMPACT project was a European Union-funded initiative to develop expertise and infrastructure for libraries digitizing the textual heritage of Europe. Despite the rapid rate of text digitization by European libraries, the availability of full-text transcriptions was not keeping pace. With many libraries solving the same digitization challenges, solutions to problems were being duplicated, leading to inefficient use of time and resources. Moreover, existing OCR software produced unsatisfactory accuracy for historical printed books. Through the formation of a pan-European consortium of libraries, the IMPACT project consolidated digitization expertise and developed tools, resources, and best practices to surmount the challenges of digitization on such an extensive scale. The project lasted from 2008- 2012. Among its achievements were the monumental creation of the IMPACT dataset of historical document images with ground truth for text and layout analysis, the development of software tools for layout analysis, ground truth creation, and optical character recognition post-correction, the proposal of the PAGE format, and the exploration of techniques for OCR, layout analysis, and image correction (Papadopoulos 2013; Pletschacher & Antonacopoulos 2010; Vobl et al 2014).

The Early Modern OCR Project (eMOP) was an effort by researchers at Texas A & M University to produce transcriptions of the Early English Books Online and 18th Century Collections Online databases. Containing nearly 45 million pages collectively, these two commercial databases are essential tools for historians studying the literature of the 15th through the 18th century. The project produced accurate transcriptions paired with the corresponding text images and made available for crowd-sourced post-correction on the 18thConnect website using the TypeWright tool; it developed a true "Big Data" infrastructure to take advantage of high-performance computing resources for both OCR and image post-processing. Another important contribution was the pioneering work on a historical font database (Heil and Samuelson 2013).

5 CONCLUSIONS

Historical Document Processing transforms scanned documents from the past into digital transcriptions for the future. After pre-processing through binarization, layout analysis, and line segmentation, the images of individual lines are converted into digital text through either HTR or OCR. Within the past decade, first conventional machine learning techniques using handcrafted features and more recently neural network-driven methodologies have become solutions to producing accurate transcriptions from historical texts from medieval manuscripts and fifteenth-century incunabula through early modern printed works. Projects such as IMPACT, Transcriptorium, eMOP, and HisDoc have made significant contributions to advancing the scholarship of the field and creating vital datasets and software tools. The combined expertise of computer scientists, digital humanists, historians, and archivists will all be necessary to meet the challenge of HDP for the future.

As archives continue to be digitized, the volume and variety of archival data and the velocity of its creation clearly indicate that this is a "Big Data" challenge. Accurate transcriptions are a prerequisite for meaningful information retrieval in archival documents. The creation of robust tools and infrastructure for this new phase of historical document processing will be the mandate of all those who wish to preserve humanity's historical textual heritage in the digital age.

ACKNOWLEDGEMENTS

This research is supported in part by REU grant #1560037 from the National Science Foundation.

REFERENCES

Baechler, M., & Ingold, R. (2010). Medieval manuscript layout model. Proceedings of the 10th ACM Symposium on Document Engineering - DocEng '10, 275.

Bamman, D., & Smith, D. (2012). Extracting two thousand years of latin from a million book library. Journal on Computing and Cultural Heritage, 5(1), 1–13.

Ben Messaoud, I., Amiri, H., El Abed, H., & Märgner, V. (2012). Binarization effects on results of text-line segmentation methods applied on historical documents. 2012 11th International Conference on Information Science, Signal Processing and Their Applications (ISSPA), 1092–1097.

Bosch, V., Toselli, A. H., & Vidal, E. (2014). Semiautomatic text baseline detection in large historical handwritten documents. 2014 14th International Conference on Frontiers in Handwriting Recognition, 690–695.

Breuel, T. M., Ul-Hasan, A., Al-Azawi, M. A., & Shafait, F. (2013). High-performance ocr for printed english and fraktur using lstm networks. 2013 12th International Conference on Document Analysis and Recognition, 683–687.

Bukhari, S. S., Kadi, A., Jouneh, M. A., Mir, F. M., & Dengel, A. (2017). Anyocr: An open-source ocr system for historical archives. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 305–310.

Bukhari, S. S., Shafait, F., & Breuel, T. M. (2012). An image based performance evaluation method for page dewarping algorithms using sift features. In M. Iwamura & F. Shafait (Eds.), Camera-Based Document Analysis and Recognition (pp. 138–149). Springer.

Chandna, S., Rindone, F., Dachsbacher, C., & Stotzka, R. (2016). Quantitative exploration of large medieval manuscripts data for the codicological research. 2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV), 20–28.

Christy, M., Gupta, A., Grumbach, E., Mandell, L., Furuta, R., & Gutierrez-Osuna, R. (2018). Mass digitization of early modern texts with optical character recognition. Journal on Computing and Cultural Heritage, 11(1), 1–25.

Clausner, C., Pletschacher, S., & Antonacopoulos, A. (2011). Aletheia—An advanced document layout and text ground-truthing system for production environments. 2011 International Conference on Document Analysis and Recognition, 48–52.

Fischer, A., Baechler, M., Garz, A., Liwicki, M., & Ingold, R. (2014). A combined system for text line extraction and handwriting recognition in historical documents. 2014 11th IAPR International Workshop on Document Analysis Systems, 71–75.

Fischer, A., Bunke, H., Naji, N., Savoy, J., Baechler, M., & Ingold, R. (2012a). The hisdoc project. Automatic analysis, recognition, and retrieval of handwritten historical documents for digital libraries.

Fischer, A., Frinken, V., Fornés, A., & Bunke, H. (2011a). Transcription alignment of Latin manuscripts using hidden Markov models. Proceedings of the 2011 Workshop on Historical Document Imaging and Processing - HIP '11, 29.

Fischer, A., Frinken, V., Fornés, A., & Bunke, H. (2011b). Transcription alignment of Latin manuscripts using hidden Markov models. Proceedings of the 2011 Workshop on Historical Document Imaging and Processing - HIP '11, 29.

Fischer, A., Indermühle, E., Bunke, H., Viehhauser, G., & Stolz, M. (2010). Ground truth creation for handwriting recognition in historical documents. Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, 3–10.

Fischer, A., Indermuhle, E., Frinken, V., & Bunke, H. (2011). Hmm-based alignment of inaccurate transcriptions for historical documents. 2011 International Conference on Document Analysis and Recognition, 53–57.

Fischer, A., Keller, A., Frinken, V., & Bunke, H. (2012). Lexicon-free handwritten word spotting using character HMMs. Pattern Recognition Letters, 33(7), 934–942.

Fischer, A., Riesen, K., & Bunke, H. (2010). Graph similarity features for hmm-based handwriting recognition in historical documents. 2010 12th International Conference on Frontiers in Handwriting Recognition, 253–258.

Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G., & Stolz, M. (2009). Automatic transcription of handwritten medieval documents. 2009 15th International Conference on Virtual Systems and Multimedia, 137–142.

Frinken, V., Fischer, A., Baumgartner, M., & Bunke, H. (2014). Keyword spotting for self-training of BLSTM NN based handwriting recognition systems. Elsevier.

Frinken, V., Fischer, A., & Martínez-Hinarejos, C.-D. (2013). Handwriting recognition in historical documents using very large vocabularies. Proceedings

of the 2nd International Workshop on Historical Document Imaging and Processing - HIP '13, 67.

Gatos, B., Louloudis, G., & Stamatopoulos, N. (2014). Segmentation of historical handwritten documents into text zones and text lines. 2014 14th International Conference on Frontiers in Handwriting Recognition, 464–469.

Granell, E., Chammas, E., Likforman-Sulem, L., Martínez-Hinarejos, C.-D., Mokbel, C., & Cîrstea, B.-I. (2018). Transcription of spanish historical handwritten documents with deep neural networks. Journal of Imaging, 4(1), 15.

Heil, J., & Samuelson, T. (2013). Book history in the early modern ocr project, or, bringing balance to the force. Journal for Early Modern Cultural Studies, 13(4), 90–103.

Jenckel, M., Bukhari, S. S., & Dengel, A. (2016). Anyocr: A sequence learning based ocr system for unlabeled historical documents. 2016 23rd International Conference on Pattern Recognition (ICPR), 4035–4040.

Kahle, P., Colutto, S., Hackl, G., & Muhlberger, G. (2017). Transkribus—A service platform for transcription, recognition and retrieval of historical documents. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 19–24.

Le Bourgeois, F., & Emptoz, H. (2007). Debora: Digital access to books of the renaissance. *International Journal of Document Analysis and Recognition (IJDAR)*, *9*(2–4), 193–221.

Mas, J., Rodriguez, J. A., Karatzas, D., Sanchez, G., & Llados, J. (2008). Histosketch: A semi-automatic annotation tool for archival documents. *2008 The Eighth IAPR International Workshop on Document Analysis Systems*, 517–524.

Meyer, E. T., & Eccles, K. (2016). *The impacts of digital collections: Early english books online & house of commons parliamentary papers* (SSRN Scholarly Paper ID 2740299). Social Science Research Network.

Papadopoulos, C., Pletschacher, S., Clausner, C., & Antonacopoulos, A. (2013). The IMPACT dataset of historical document images. *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing - HIP '13*, 123.

Pintus, R., Yang, Y., & Rushmeier, H. (2015). Athena: Automatic text height extraction for the analysis of text lines in old handwritten manuscripts. *Journal on Computing and Cultural Heritage*, *8*(1), 1–25.

Pletschacher, S., & Antonacopoulos, A. (2010). The page (Page analysis and ground-truth elements) format framework. *2010 20th International Conference on Pattern Recognition*, 257–260.

Raha, P., & Chanda, B. (2019). Restoration of historical document images using convolutional neural networks. 2019 IEEE Region 10 Symposium (TENSYMP), 56–61.

Rahnemoonfar, M., & Plale, B. (2013). Automatic performance evaluation of dewarping methods in large scale digitization of historical documents. *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL '13*, 331.

Rath, T. M., & Manmatha, R. (2007). Word spotting for historical documents. *International Journal of Document Analysis and Recognition (IJDAR)*, *9*(2), 139–152.

Roe, E., & Mello, C. A. B. (2013). Binarization of color historical document images using local image equalization and xdog. *2013 12th International Conference on Document Analysis and Recognition*, 205–209.

Rydberg-Cox, J. A. (2009). Digitizing latin incunabula: Challenges, methods, and possibilities. *Digital Humanities Quarterly*, *003*(1).

Sastry, P. N., & Krishnan, R. (2012). A data acquisition and analysis system for palm leaf documents in Telugu. *Proceeding of the Workshop on Document Analysis and Recognition*, 139–146.

Serrano, N., Castro, F., & Juan, A. (2010, May). The rodrigo database. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. LREC 2010, Valletta, Malta.

Shafait, F. (2009). Document image analysis with OCRopus. *2009 IEEE 13th International Multitopic Conference*, 1–6.

Simistira, F., Seuret, M., Eichenberger, N., Garz, A., Liwicki, M., & Ingold, R. (2016). Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts. *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 471–476.

Springmann, U., & Lüdeling, A. (2017). OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus. *Digital Humanities Quarterly*, *011*(2).

Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A., & Fink, F. (2014). OCR of historical printings of Latin texts: Problems, prospects, progress. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 71–75.

Su, B., Lu, S., & Tan, C. L. (2010). Binarization of historical document images using the local maximum and minimum. *Proceedings of the 8th IAPR International Workshop on Document Analysis Systems - DAS '10*, 159–166.

Tabrizi, M. H. N. (2008). Digital archiving and data mining of historic document. *2008 International Conference on Advanced Computer Theory and Engineering*, 19–23.

Ul-Hasan, A., Bukhari, S. S., & Dengel, A. (2016). Ocroract: A sequence learning ocr system trained on isolated characters. *2016 12th IAPR Workshop on Document Analysis Systems (DAS)*, 174–179.

Vobl, T., Gotscharek, A., Reffle, U., Ringlstetter, C., & Schulz, K. U. (2014). PoCoTo—An open source system for efficient interactive postcorrection of OCRed historical texts. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 57–61.

Wei, H., Chen, K., Nicolaou, A., Liwicki, M., & Ingold, R. (2014). Investigation of feature selection for historical

document layout analysis. *2014 4th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1–6.

Würsch, M., Ingold, R., & Liwicki, M. (2016). Divaservices—A restful web service for document image analysis methods. *Digital Scholarship in the Humanities*, fqw051.

Yang, Y., Pintus, R., Gobbetti, E., & Rushmeier, H. (2017). Automatic single page-based algorithms for medieval manuscript analysis. *Journal on Computing and Cultural Heritage*, *10*(2), 1–22.