HADARA – A Software System for Semi-Automatic Processing of Historical Handwritten Arabic Documents

Werner Pantke, Volker Märgner, Daniel Fecker, Tim Fingscheidt; Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany
Abedelkadir Asi, Ofer Biller, Jihad El-Sana; Ben-Gurion University of the Negev, Be'er-Sheva, Israel
Raid Saabni; Faculty of Engineering, Tel-Aviv University and Triangle R&D Center, Kafr Qara, Israel
Mohammad Yehia; Triangle R&D Center, Kafr Qara, Israel

Abstract

Recently, many big libraries all over the world have been scanning their collections to make them publicly available and to preserve historical documents. We present a modular software system which can be used as a tool for semi-automatical processing of historical handwritten Arabic documents. The development of this system is part of the HADARA project which aims for historical document analysis of Arabic manuscripts and consists of a project team including engineers and computer scientists but also users such as linguists and historians. The HADARA system is designed to support script and content analysis, identification, and classification of historical Arabic documents. The system has been created following an iterative development approach, and the current version assists the user in an interactive and partially already in an automatic manner. In this paper, a system overview is given and the first modules are presented which support the annotation of a scanned manuscript in a semi-automatic manner. They comprise page layout analysis, text line segmentation, and transcription. Word spotting is the first application implemented in the HADARA system and its concept is outlined in this paper.

Introduction

Nowadays, there is a trend to digitize printed or handwritten historical documents in many big libraries all over the world. Scanned images are published on library websites after manually adding metadata information. This information provides the ability to search for a specific document in a large database. Unfortunately, searching through the content of a document is not possible as long as the content itself is not digitally available in a textual form. However, a manual transcription requires large efforts in terms of time and costs. To overcome these limitations computer scientists and researchers in the field of document analysis and recognition develop algoritms for an automatic transcription of scanned historical documents or at least to support specific parts of this task. Pattern recognition methods, such as automatic text recognition or word spotting, are employed for this purpose.

On the one hand, large scale digitization projects are underway at most big libraries and even private companies (e. g., Million Book Project [1] or Google Book Search [2]) to preserve paper documents in digital format. On the other hand, many researchers all over the world work on projects for historical document processing and recognition, as can be seen at conferences like ICDAR [3] , ICFHR [4], or DAS [5], workshops like HIP [6], and research projects like IMPACT [7]. The cooperation of experts from digitization projects and researchers from the field of document analysis and recognition is gaining importance. For example, the very time consuming and expensive task of the generation of training data for recognition tasks can benefit from cooperation of librarians and computer scientists.

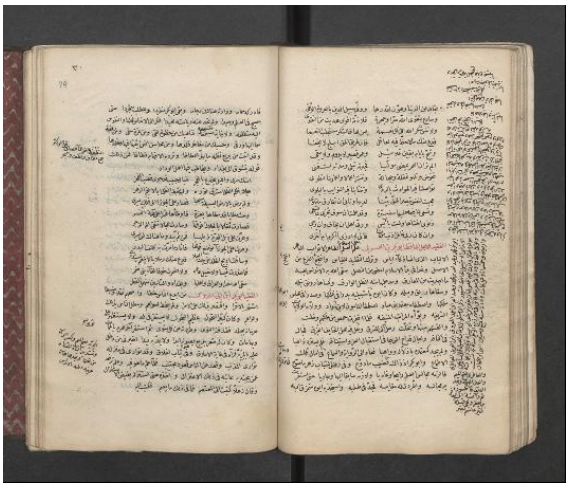


Figure 1. *Example pages of a scanned Arabic handwritten book with side notes*[1]

Typical pages of an Arabic handwritten book written in the 18th century (the text dates back to the 12th century) are shown in Figure 1. In the center of each page the main body text is written, additionally comments and remarks are written on the page borders. These are added usually neither from the same writer nor from the same period as the main body text. This example shows one of the major problems for interpretation of historical Arabic manuscripts. An overview about the challenges of historical document processing is given in [8].

The HADARA project team consists of scientists from signal processing, computer science, science of history, and linguistics. The diversity inside this group ensures that the different needs of the involved areas of expertise are respected during the whole development process. The core of the HADARA system consists of an easy-to-use historical document processing tool chain

[1] Scan provided by the Damascene family library Refaiya at the University Library in Leipzig, Germany, Website: http://www.refaiya.uni-leipzig.de/

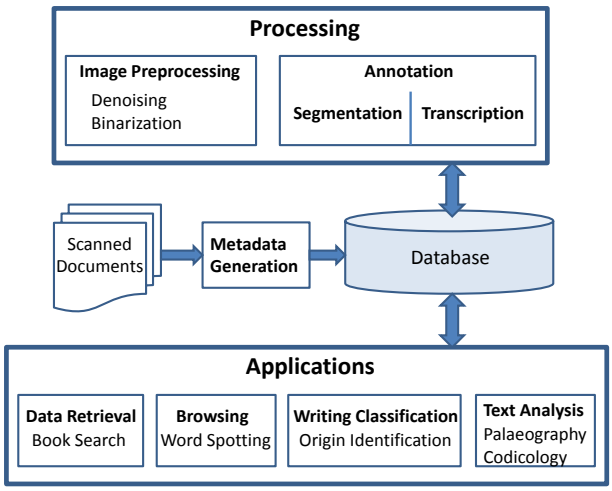


Figure 2. *Overview of the HADARA system*

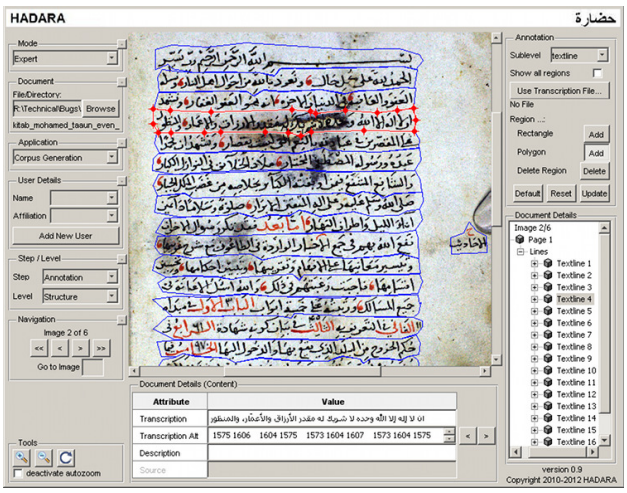


Figure 3. *Graphical user interface of the HADARA system showing the annotation step*

that allows to process digitized historical documents, and to perform search queries on them. The system is not only intended to be used by librarians, e. g., to assign metadata to scanned manuscripts or to annotate and transcribe the content, but also by historians to systematically process and index historical Arabic manuscripts. Besides dealing with handwritten documents and Arabic language, a further characteristic of the HADARA system is its mobility and compactness. This characteristic is accomplished by using a mobile scanning device [9] for digitizing rare books in small private libraries and implementing an easy-to-use graphical user interface (GUI) which is user-friendly and easy to handle after a short training period.

Furthermore, it is important to mention that the presented system can generally be used for any language and alphabet, for handwritten as well as printed text. Indeed, our system targets handwritten Arabic documents, revealing specifics such as connected characters, varying classes of dots, and diacritical signs. However, it is also able to handle left-to-right languages and alphabets that do not need these special treatments, like Latin character alphabets.

The paper is organized as follows: The next section gives an overview of the HADARA system. Some modules supporting the important annotation step are then described in the subsequent sections: automatic page layout analysis, text line segmentation, and semi-automatic transcription. Consecutively, word spotting is discussed as a first application that makes use of the HADARA system framework. Finally, the paper ends with concluding remarks.

System Overview

The block diagram of Figure 2 gives an overview of the HADARA system. Scanned documents, in our case Arabic manuscripts, are manually attached with corresponding metadata and stored in a database. These data are used and modified by the main components of the system, divided into the two blocks *Processing* and *Applications*. In the following, some aspects of the system are discussed in detail.

A first important component of the system is the data format used to store all data of each document during different levels of processing. To allow both flexibility and the possibility to exchange data easily, we use an XML format based on standards used by many groups in this field. The metadata, for example, are stored in a format derived from EUROPEANA [10] and Dublin Core [11], but extended to deal with typical information of Arabic manuscripts.

The processing block consists of several processing steps, e. g., image preprocessing and annotation. Each step can be done manually or automatically and is realized by one or more modules. The first step handles the preprocessing of page images with modules such as noise reduction and binarization. The preprocessing step is required by most of the automatic processing steps. The second processing step is the annotation comprising segmentation and transcription modules. A basic segmentation is realized by a page layout analysis (PLA) module that extracts the logical structure of a page followed by a text line and word segmentation module that extracts lines and words out of text block segments. This segmentation is a necessary as additional preprocessing for most subsequent automatic applications. Transcription modules provide a symbolic representation of text shown on the scanned images. The automatic transcription of scanned manuscripts is called handwriting recognition.

The applications block shows different applications already implemented or planned to be implemented in the HADARA system. The simplest application is data retrieval using the metadata linked to any scanned manuscript to search for a specific document and show the corresponding scans to the user. A word spotting application supports researchers seeking for specific words in a document even without a transcription available. The origin identification is another application researchers are interested in with the goal to identify the writer, region, or time a manuscript originates. The most challenging applications are codicology and paleography which need text images and corresponding transcriptions. These applications comprise the study of features like marginalia, glosses, ownership inscriptions, and aspects of decoration.

The annotation step and all applications may run with different levels of automation. The goal of the HADARA project is to

implement step by step more and more modules that support automatic processing. To use the system during the processing with different levels of automation, the graphical user interface (GUI) plays an important role. Figure 3 shows a screen shot of the GUI during the interactive annotation step. A page image is shown, in which segmented text lines can be selected and transcribed using the input fields at the bottom. In case of Arabic text, an automatically converted ASCII representation is displayed as alternative transcription readable to users who are not capable of reading Arabic characters. Left-to-right and other right-to-left languages are also supported by the input field for the transcription.

With the help of the GUI most processing steps may be done interactively or automatically. The verification of processing steps and interactive testing of automatic modules are additional tools provided by the GUI to support researchers working with historical documents. The more applications are supported by automatic modules like text recognition or word spotting the easier and faster a researcher's task can be addressed.

In the following, we present in some detail modules of automatic or semi-automatic approaches for page layout analysis, text line segmentation, transcription, and word spotting.

Processing: Page Layout Analysis

Figure 1 shows one of the typical problems of historical Arabic manuscript processing: Many text blocks in different orientation are written on the same page. Automatic processing of such a page needs the separation of different text blocks from each other. This is one of the tasks of page layout analysis. Different approaches are published to solve this problem. Some try to extract each single text block of a document page [12], while others try to separate the main body text from all the additional text and other content located on page margins. A method based on the latter approach is presented below.

Scholars tended to add notes on page margins mainly because paper was an expensive material (see Figure 1). These notes, which are also known as side notes, might contain important information for professionals and they were not written by the same writer of the main body text. Extracting side notes is a step of great importance for other steps in the segmentation pipeline, e. g., text-line extraction algorithms which assume clean page margins [13, 14] and, therefore, require extracted text blocks. We suggest a connected component-based method that reliably classifies each component to one of the two text classes, main body or side note.

Our method trains a classifier to distinguish between the classes by exploiting simple features, yet representative and distinguishable, from connected components. As it is already known, parameter tuning is a real challenge for machine learning techniques. In this work we use the AutoMLP classifier [15] which is a self-tunable neural network to adjust both the learning rate and the number of nodes in the hidden layer. The generation of the classifier combines ideas from genetic algorithms and stochastic optimization. It maintains a small ensemble of networks that are trained in parallel with different learning rates and different numbers of hidden nodes. After a small number of iterations, the error rate is determined on an internal validation set and the worst performers are replaced with copies of the best networks, modified to have different numbers of hidden nodes and learning rates. This process is repeated until a specific threshold on the error rate is reached.

(a) (b)

Figure 4. *In dark letters: (a) Coarse classification (b) final classification after the refinement step*

We have divided our features into two main categories, component shape and component context features. We use the raw shape of components as one of the main features. Additional characteristics of an individual component are also used such as height, area, relative distance and orientation. The orientation of a connected component is calculated with respect to its neighborhood. We apply a projection profile on the considered neighborhood in 12 directions. A robustness measure is computed for each profile and the angle that corresponds to the most robust profile is chosen as the connected component orientation. The robustness is computed as

$$s = \frac{1}{N}\sum_{n=1}^{N}(y_h^{(n)} - y_l^{(n)}), \tag{1}$$

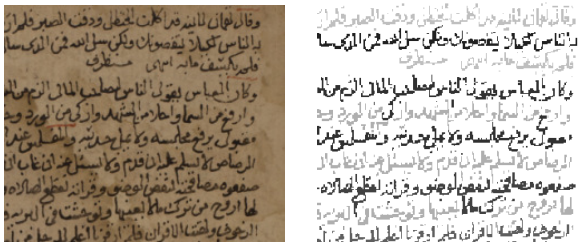
where N is the number of peaks found in the profile, $y_h^{(n)}$ is the value of the n-th peak, and $y_l^{(n)}$ is the value of the highest valley around the n-th peak. In addition, we exploit the neighborhood of each component as we believe that it incorporates important information about density and regularity of text. To further improve the reliability of the context features, one can compute a precise neighborhood dimension using evolution maps as in [16].

At this level of the algorithm the classifier produces a coarse segmentation (see Figure 4(a)). We refine the classification of each component by employing weighted nearest neighbor (NN) analysis for its neighborhood (see Figure 4(b)). Several neighborhood dimensions were examined in order to determine the optimal dimensions. It is important to emphasize that we do not calculate new weights, but we rather use the class probabilities of each component which were already computed during the coarse classification phase.

For evaluation purposes we have manually generated a pixel-based ground truth. We calculate the classification accuracy for main body text and side notes separately (see Table 1). By combining precision and recall values, the F-measure depicts the classification accuracy adequately. Therefore, we adopt the F-measure in the evaluation process. For side notes, a classification accuracy of 95 % has been achieved. The refining step improves the algorithm accuracy, but it might fail in border regions where both main body text and side notes overlap or touch one

Table 1. Performance evaluation of our page layout analysis method for both main body and side note text for different post-processing window sizes

Window Size	Main Body F-Measure (%)	Side Notes F-Measure (%)
50	91.37	90.74
100	94.34	93.93
150	**95.02**	**94.68**
200	94.65	94.22
250	93.91	93.35



(a) (b)

Figure 5. *Example text line segmentation with (a) showing a part of a scanned page image and (b) the respective segmented text lines*

another. For further details considering different parts of the algorithm please refer to the original paper [17].

Processing: Text Line Segmentation

The segmentation of a text block into text lines is another required step towards automatic text recognition. Text line segmentation in many cases is a simple task easy to be performed using projection methods [18]. These methods often fail particularly for historical handwritten documents due to background noise (e. g., caused by paper aging) and skewed curved text lines [19] as can be seen in Figure 5. Additionally, in case of Arabic manuscripts we have to deal with segmentation problems caused by diacritics, overlapping characters, connecting text lines, and different character sizes even in text blocks on the same page.

To solve some of these problems we have integrated different text line segmentation modules into the HADARA system [13]. A very promising approach is based on connected components. To get an image with connected components of the handwritten text an optimized binarization approach (e. g., [20], [21], [22]) followed by a subsequent connected components detection is used. Instead of processing all black pixels of a binary image, only the centers of gravity of each connected component are taken into account. The 10 % smallest and largest components, respectively, are disregarded while a projection of the centers of gravity on the vertical axis is applied. The resulting projection profile is low-pass filtered. Maxima of this profile are assumed to represent centers of text lines. Finally, all connected components of the text block are assigned to the line to which they have the lowest vertical distance.

In Figure 5, segmented text lines are illustrated using alternating colors. Even in overlapping and slightly curved lines this approach detects the correct line boundaries and assigns most of the diacritics to the correct line. Errors still happen especially in case of touching objects from neighboring lines. The results of the text line segmentation can be used as a basis for a more granular segmentation into words, graphemes, and characters if needed.

Processing: Semi-Automatic Transcription

Text block and text line segmentation are prerequisites for the transcription task. The transcription is performed line by line. Doing it manually is a very time consuming and error-prone task. Automatic Arabic handwriting recognition is a research field today which can be very helpful in supporting the transcription process. Many approaches have been published recently for handwritten Arabic word or even text recognition [23]. But still a universal recognizer, especially for historical manuscripts, does not exist. Main problems are the knowledge of the vocabulary (words that appear in the text) and of the specific writing styles used in the manuscript to recognize. Therefore, the development of a recognizer for historical Arabic manuscripts is an important task within the HADARA project.

During the transcription process, textual transcriptions are assigned to previously segmented text lines, words, or other entities. Corresponding pairs of word images and word transcriptions, for example, can then be used by historians to analyze the text and also to train or improve a recognizer for automatic text recognition. If the transcription of a document is already available but the corresponding links to word locations in the document images are missing, a transcription alignment as in [24] could be applied. Our focus, however, are documents without available transcription.

To support the time-consuming transcription task, a strategy for a semi-automatic transcription is shown in the block diagram in Figure 6. After the scanning and preprocessing of a document and if no previously trained recognizer is available, the process begins with a manual transcription of a certain number of page images. For this task, the HADARA system offers a tool which guides through the transcription process, providing access to both manual and semi-automatic modules. Additionally, there are automatic segmentation modules available to support this work. As soon as a relevant amount of page images has manually been processed and verified, the first training of a recognizer that employs, e. g., hidden Markov models (HMMs) can be accomplished based on this data. A next set of scanned document pages can then be transcribed automatically using the trained recognizer, also followed by a manual verification. To iteratively increase the recognition accuracy, the recognizer is retrained using the extended data set. This process continues until all page images are processed and a fully annotated document is obtained. If a recognizer for a specific font or writing style is already available at the beginning of the transcription process, it can be employed to replace the otherwise required manual transcription.

Application: Word Spotting

As mentioned before, handwritten text recognition especially of historical Arabic manuscripts is a research task that is still widely unresolved today. Therefore, word spotting instead of complete recognition of text has been proposed as an alternative. Word spotting means the spotting of single words in document images without knowledge of the underlying text. This technique may be used for browsing, searching, or semi-automatic indexing of scanned historical manuscripts.

Using word spotting, researchers can work on manuscripts searching for certain words even without preceding segmentation

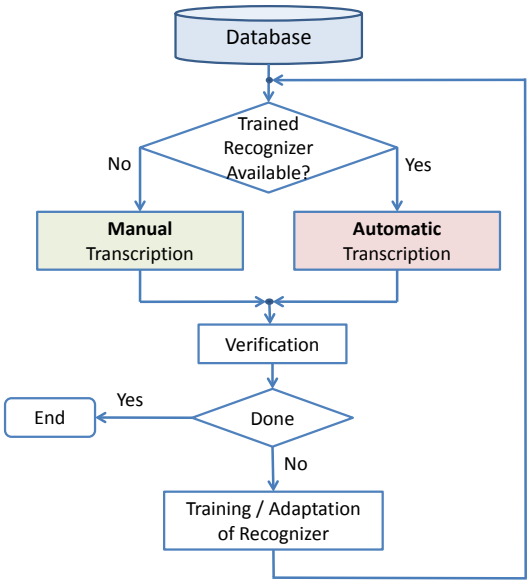


Figure 6. *A semi-automatic transcription process*

and transcription steps. Different approaches have been developed during the last years. Learning-based methods that support searching for a textual keyword can be used if enough training data are available. In contrast, template-based methods require an example image of the word being searched for, the template. No prior training is needed but at least one template has to be found and selected from the manuscript under investigation.

Another application of word spotting in historical documents is grouping similar word images into separate clusters. Once such a clustering of a manuscript exists, the number of words contained in a cluster can be used as a cue for determining the importance of the word in a document. For this purpose, highly frequent terms of a language, such as ”the”, ”a”, ”an”, ”and”, ”of” for English, are so-called stop words and discarded. All interesting clusters with terms that are deemed important can then be manually annotated which makes it possible to construct a partial index, which links words to the locations where they occur. In case of historical Arabic manuscripts with its cursive writing style, a word segmentation is usually not satisfying. Instead, approaches using sliding windows on whole text lines can be employed.

Into the HADARA system, some template-based approaches have been integrated [25, 26, 27, 28]. First tests show encouraging results on some typical Arabic manuscripts.

Conclusions

We presented a software system to process historical handwritten Arabic documents with its data representation and user interface being core components. The data format is based on public standards and the user interface is developed together with users to meet their expectations. We presented approaches for page layout analysis and text line segmentation as important steps for the automation of the annotation process. The concept of semi-automatic transcription completes our system, on which applications can be built atop. Finally, word spotting is discussed as example application that can be used for research on historical Arabic documents.

The presented system is intended for being used by small groups or libraries to support research projects on historical documents and in parallel collecting and annotating data to be used for the training of recognition tasks. Next steps in this project are the integration of a recognition module and the further optimization of the system, its modules, and its user interface. We plan to make the system available to the public eventually.

Acknowledgments

We would like to acknowledge the contribution of the historical linguist Karim Salman in the field of Arabic document selection, scanning, and annotation. But also his comments to the usage of the GUI have always been very beneficial during the development process. The German Research Foundation is gratefully acknowledged for support of this project under contract FI 1494/3-2.

References

[1] The Universal Library (The Million Book Project). [Online]. Available: http://archive.org/details/millionbooks

[2] Google Books. [Online]. Available: http://books.google.com/

[3] *International Conference on Document Analysis and Recognition*. Beijing, China: IEEE, 2011.

[4] *International Conference on Frontiers in Handwriting Recognition*. Bari, Italy: IEEE, 2012.

[5] M. Blumenstein, U. Pal, and S. Uchida, Eds., *International Workshop on Document Analysis Systems*, IAPR. Gold Coast, Queensland, Australia: IEEE, 2012.

[6] B. Barrett, M. S. Brown, R. Manmatha, and J. Gehring, Eds., *1st Workshop on Historical Document Imaging and Processing*. Beijing, China: ACM, 2011.

[7] Improving Access to Text (IMPACT). [Online]. Available: http://www.impact-project.eu/

[8] A. Antonacopoulos and A. C. Downton, “Special issue on the analysis of historical documents,” *International Journal on Document Analysis and Recognition*, vol. 9, no. 2-4, pp. 75–77, 2007.

[9] Traveller’s Concersation Copy Stand, TCCS 4232, Vestigia, University Library Graz, Austria. [Online]. Available: http://www.vestigia.at/der_traveller.html

[10] EUROPEANA. EUROPEANA Foundation. [Online]. Available: http://pro.europeana.eu/edm-documentation

[11] Dublin Core Metadata Initiative. [Online]. Available: http://dublincore.org/

[12] N. Ouwayed and A. Belaïd, “A general approach for multi-oriented text line extraction of handwritten documents,” *International Journal on Document Analysis and Recognition*, vol. 15, no. 4, pp. 297–314, 2012.

[13] A. Asi, R. Saabni, and J. El-Sana, “Text line segmentation for gray scale historical document images,” in *Proceedings of the Workshop on Historical Document Imaging and Processing*, Beijing, China, 2011, pp. 120–126.

[14] I. B. Yosef, N. Hagbi, K. Kedem, and I. Dinstein, “Text line segmentation for degraded handwritten historical documents,” in *Proceedings of the International Conference on Document Analysis and Recognition*, Barcelona, Spain, 2009, pp. 1161–1165.

[15] T. Breuel and F. Shafait, “AutoMLP: Simple, effective, fully automated learning rate and size adjustment,” in *The Learning Workshop*, Snowbird, Utah, USA, 2010.

[16] O. Biller, K. Kedem, I. Dinstein, and J. El-Sana, “Evolution maps

for connected components in text documents," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, Bari, Italy, 2012, pp. 405–410.

[17] S. S. Bukhari, A. Asi, T. Breuel, and J. El-Sana, "Layout analysis for arabic historical document images using machine learning," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, Bari, Italy, 2012.

[18] R. P. d. Santos, G. S. Clemente, T. I. Ren, and G. D. C. Cavalcanti, "Text line segmentation based on morphology and histogram projection," in *Proceedings of the International Conference on Document Analysis and Recognition*. Barcelona, Spain: IEEE Computer Society, 2009, pp. 651–655. [Online]. Available: http://dx.doi.org/10.1109/ICDAR.2009.183

[19] L. Likforman-Sulem, A. Zahour, and B. Taconet, "Text line segmentation of historical documents: a survey," *International Journal on Document Analysis and Recognition*, vol. 9, no. 2, pp. 123–138, 2007. [Online]. Available: http://dx.doi.org/10.1007/s10032-006-0023-z

[20] I. B. Messaoud, H. Amiri, H. El Abed, and V. Märgner, "New binarization approach based on text block extraction," in *Proceedings of the International Conference on Document Analysis and Recognition*, Beijing, China, 2011, pp. 1205–1209.

[21] S. Lu, B. Su, and C. L. Tan, "Document image binarization using background estimation and stroke edges," *International Journal on Document Analysis and Recognition*, vol. 13, no. 4, pp. 303–314, Dec. 2010. [Online]. Available: http://dx.doi.org/10.1007/s10032-010-0130-8

[22] I. Bar-Yosef, I. Beckman, K. Kedem, and I. Dinstein, "Binarization, character extraction, and writer identification of historical hebrew calligraphy documents," *International Journal on Document Analysis and Recognition*, vol. 9, no. 2-4, pp. 89–99, 2007.

[23] V. Märgner and H. El Abed, Eds., *Guide to OCR for Arabic Scripts*, ser. Advances in Pattern Recognition. Springer Verlag, 2012.

[24] A. Fischer, V. Frinken, A. Fornés, and H. Bunke, "Transcription alignment of latin manuscripts using hidden markov models," in *Proceedings of the Workshop on Historical Document Imaging and Processing*. Beijing, China: ACM, 2011, pp. 29–36. [Online]. Available: http://doi.acm.org/10.1145/2037342.2037348

[25] Y. Leydier, F. Lebourgeois, and H. Emptoz, "Text search for medieval manuscript images," *Pattern Recognition*, vol. 40, no. 12, pp. 3552–3567, 2007.

[26] Y. Leydier, A. Ouji, F. Lebourgeois, and H. Emptoz, "Towards an omnilingual word retrieval system for ancient manuscripts," *Pattern Recognition*, vol. 42, no. 9, pp. 2089–2105, Apr. 2009.

[27] R. Saabni and J. El-Sana, "Keyword searching for arabic handwritten documents," in *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, Montreal, Canada, 2008, pp. 716–722.

[28] ——, "Word spotting for handwritten documents using chamfer distance and dynamic time warping," in *Proceedings of the Document Recognition and Retrieval Conference, part of the IS&T-SPIE Electronic Imaging Symposium*, ser. SPIE Proceedings, vol. 7874. San Jose, CA, USA: SPIE, 2011, pp. 1–10.

Author Biography

Werner Pantke received his diploma in computer science from Technische Universität Braunschweig (TU-BS), Braunschweig, Germany. Since 2010, he is both a PhD student and working as a researcher at the Institute for Communications Technology of TU-BS. His main area of research is word spotting and handwriting recognition, especially for historical Arabic manuscripts.

Volker Märgner received his diploma (Dipl.-Ing.) and doctorate (Dr.-Ing.) degrees in electrical engineering from Technische Universität Braunschweig (TU-BS), Germany, in 1974 and 1983 respectively. Since 1983, he has been working at TU-BS where he currently is a member of the research and teaching staff at the Institute for Communications Technology. His main research interests are pattern recognition and historic document image analysis

Daniel Fecker received his diploma in computer and communications systems engineering from Technische Universität Braunschweig (TU-BS), Braunschweig, Germany. Since 2009, he is both a PhD student and working as a researcher at the Institute for Communications Technology of TU-BS. His main area of research is training of classifiers with highly imbalanced datasets for industrial quality control and writer identification.

Tim Fingscheidt received the Dipl.-Ing. and the Dr.-Ing. degrees from RWTH Aachen University, Germany. From 1998 he worked with AT&T Labs, Florham Park, NJ, USA. In 1999 he joined Siemens AG (COM Mobile Devices) in Munich, Germany, from 2001 as team leader for Audio Applications. In 2005 he joined Siemens Corporate Technology in Munich, leading the company's speech technology development activities. Since 2006 he is Professor at the Institute for Communications Technology at Technische Universität Braunschweig, Germany. His research interests are speech and audio signal processing and pattern recognition.

Abedelkadir Asi is a PhD student in the computer science department at Ben-Gurion University of the Negev. He recieved his B.Sc. from the Hebrew University of Jerusalem and his M.Sc. from Ben-Gurion University, both in computer science. His research interest is focused on historical document image analysis.

Ofer Biller is a PhD student in computer science at Ben-Gurion University of the Negev (BGU). He earned his B.Sc. from the Technion institute, Israel and his M.Sc. from BGU. He has been involved in the industry where he served as a technical leader in a software company. His research interest is in the field of historical documents analysis.

Jihad El-Sana is an Associated Professor in the Department of Computer Science. He received his B.Sc. and M.Sc. in Computer Science from BGU. In 1995 he won a Fulbright Scholarship for Israeli Arabs, for doctoral studies in the US. In 1999 he earned a PhD in Computer Science from the State University of New York, Stony Brook under the supervision of Amtabh Varshney. His research focuses on developing graphics technologies that simplify the generation of augmented reality system. He has made a significant contribution in developing processing tools that recognize online Arabic handwriting and the revision of Arabic historical documents.

Raid Saabni is a senior researcher at the triangle Research & Development center and a lecturer at the Tel Aviv Yafo Academic College. He received his B.Sc. in Mathematics and Computer Science in 1989 and his M.Sc. and PhD in Computer Science from Ben-Gurion University in the Negev in 2006 and 2010 respectively. His research interest is historical document image analysis, handwriting recognition, image retrieval and image processing.

Mohammad Yahia received his MD (Dr. med.) degree in medicine from Johan Wolfgang Goethe University in Frankfurt/M, Germany in 1985. From 1977 to 1984 he has been studying and researching history of medieval Arabic islamic-sciences in the institute for the history of natural sciences in Goethe University. Among his academic interests, he is researching the field of re-reading and revising the Arabic/islamic scientific heritage using modern technologies and scientific tools.