ARCHEOLOGIE TRA ORIENTE E OCCIDENTE 1 (2022)

Arabic-Character Historical Document Processing: Why and How To?

Zied Mnasri
University of Napoli L'Orientale

Abstract

The aim of Arabic-character Historical Document Processing (HDP) is to design and develop techniques that will enable automatic transcription into text files, such as in .txt or .doc format, of historic manuscripts in Arabic characters, not only for Arabic, but also for other languages based on this character, such as Farsi, Urdu, Azari, ottoman Turkish, etc. The key idea is to go from the scanned image of the manuscript to the text file using artificial intelligence techniques to accomplish two main steps: First, processing the manuscript image to identify the characters and to remove other forms generally found in historic manuscripts, such as images and other types of ornaments; secondly, identifying the characters by pattern recognition. Such a work requires the availability of a rich dataset of Arabic-character manuscripts, in addition to effective methods for image processing, pattern recognition and, optionally, language modelling. In this paper, an overview of the Arabic-character HDP state of the art, datasets, challenges, methods and potential applications is presented, as a first step to set a general framework to undertake such a project.

Keywords: Arabic manuscript, Historical Document Processing (HDP), Optical Character Recognition (OCR), character segmentation and recognition
Citation: Mnasri, Z. (2022) Arabic-Character Historical Document Processing: Why and How to? *Archeologie tra Oriente e Occidente* 1, 47-59. https://doi.org/10.6093/archeologie/9857
Corresponding author: zmnasri@unior.it

The process of digitizing old written documents so that historians and other researchers can utilize them in the future is known as historical document processing (HDP). In order to automatically transform images of old manuscripts, letters, diaries, and early printed texts into a digital format usable in data mining and information retrieval systems, it incorporates algorithms and software tools from a variety of computer science subfields, including computer vision, document analysis and recognition, natural language processing, and machine learning. The necessity to transcribe the complete text from these collections has become urgent over the past twenty years as libraries, museums, and other cultural heritage organizations have scanned an increasing amount of their historical document archives.

The processing of a Historical Arabic Documents is a difficult task. The challenges can be grouped into three main categories: First, due to the nature of Arabic script, since the letters of Arabic-character text are connected together along a writing line, secondly, the multiplicity of fonts in old Arabic-character manuscripts, that depend on the historic period and the geographic area, and thirdly, due to the fact that an important part of the relevant Arabic-character documents dates from more than ten centuries.

In addition to the main output, i.e. a text file containing the transcription in digitized characters, applying HDP for Arabic-character manuscripts may yield additional outputs, mainly:

- The possibility of adding a speech synthesis component that allows the manuscript to be read aloud from the transcribed text using speech synthesis tools, thus providing a better way of utilizing manuscripts;
- Functions that enable the creation of manuscript indexes, keyword extraction, lexical recurrence search, etc.;

- Identify the language of the manuscript, e.g. Arabic, Farsi, Urdu, etc. or detect the existence of more than a language in the same manuscript.

Thus, throughout this paper, the main aspects of Arabic-character HDP are went through, to provide an overview of the whole topic. The rest of this article is organized as follows: in section 2, the state of the art is reviewed, then in section 3, the main challenges are presented. Datasets are described in section 4, whereas methods and techniques are detailed in section 5, and potential applications are proposed in section 6. Finally, a summary is presented in the conclusion along with some ending remarks.

State of the art

Thanks to the development of document processing techniques in the last two decades, such as document image analysis (DIA), optical character recognition (OCR) and handwritten text recognition (HTR), several software solutions have been proposed in the literature for HDP, mainly for Latin-character manuscripts, and also for Arabic-character ones.

A complete set of tools for preprocessing, machine learning training, and transcription are offered in two main software systems. One of them is DIVA-Services (Würsch 2017) and the other is the Transkribus platform (Kahle 2017) from the EU-sponsored READ project (READ Project s.d.).

A web-based service called DIVA-Services offers an API (application programable interface) of tools for each level of processing historical documents. Common document image analysis (DIA) and machine learning techniques are available, according to (Würsch 2017) using a web-based interface or an API. It eases the workload for specialists in cultural heritage and computer science. A group of HDP services for DIA, HTR, and OCR are included in DIVA-Services. It is more appropriate for projects requiring an eclectic approach to tools in a tailored software toolchain or for research use-cases. There are no fees associated with commercial software licensing because it is completely open source.

The Transkribus software (Kahle 2017) platform aims to offer a complete HDP toolchain to archival professionals, cultural heritage researchers, and computer scientists. Users upload document picture files and already created layout or transcription data using the desktop program client. The client gives users the option of manually or automatically segmenting document pictures for layout analysis. Additionally, authorized users of the service have access to fresh HTR and OCR datasets for machine learning model training. A distinct RNN-based tool from the University of Valencia's Pattern Recognition and Human Language Technology research group is used for HTR whereas the ABBY Fine Reader SDK is used for OCR. The program will assist crowd-sourcing transcribing projects through the anticipated deployment of a web-based interface to supplement the desktop client.

Some software is suggested in the literature to handle historical Arabic documents with the main goal of making use of the materials easier. The majority of this software is proprietary and not accessible to non-academic public. Due to the lack of tools, it is difficult to verify their effectiveness using latest handwritten Arabic datasets. In order to create ground truth, many researchers utilize the free annotation program WebGT (Khedher 2020).

WebGT is an interactive Web-based system (Biller 2013), that can be utilized to aggregate and annotate lower hierarchy items to build higher hierarchy elements. The two file formats that the system supports are XML and CSV. It accepts scanned documents as input and produces an XML file with annotation data for each word. In this tool, a framework to make it easier to interpret old Arabic writings with complicated layouts is proposed. It contains unwarping, categorization, and localization methods for text. For each assignment, authors include already known algorithms. The side-text is separated from the main text using the coarse-fine method of Asi (2014).

In Khader (2014), an interactive annotation tool is proposed. It takes as input a scanned document and outputs XML file that contains the annotation information for the respective words. The system enables the user to store the annotated text and the corresponding word locations in XML files. For the binarization task, authors integrate existing algorithms such as the Otsu (1979) algorithm.

Regarding image processing software, a tool is proposed in Boussellaa (2007). It includes preprocessing and analysis tasks. As preprocessing tasks, it integrates binarization, skew correction (Srihari 1989) and Background/Foreground separation. Concerning analysis tasks, it includes text/graphic segmentation (Boussellaa 2006) and text line segmentation (Zahour 2004). In both cases, authors mostly integrate existing approaches. The tool is not available for public use. For line detection, the approach proposed in Cohen (2015) is integrated into the tool. For text unwarping, a simple approach based on affine transformations is proposed. As searching tool, a search engine of ancient Arabic manuscripts based on meta-data and XML annotations is proposed (but not yet available).

A software for Optical Character Recognition (OCR) is proposed in Stahlberg (2016) and includes the kaldi-based recognizer approach (Povey 2011). The tool takes as input documents images and output XML files contains transcriptions and layout information.

Challenges

The main challenges facing Arabic-character HDP are related to two main obstacles: the Arabic script morphology and the state of historical documents in Arabic character (Khedher 2020).

Analyzing Arabic script is difficult for a number of reasons.

Challenges due to Arabic script:

- Arabic text is written with the letters connected along the writing line;
- Any document analysis system should take into consideration the presence of diacritic signs that eventually denote short vowels or other pronunciation modes in Arabic text. These marks can modify the meaning of a word and are found in Arabic text as dots, punctuation below or above letters, etc.;
- Depending on the position in the word, the form of the letter varies. The same letter might look radically different at the start and end of a word. This increased the alphabet's size from 29 to nearly 160 different form of letters when all variants are taken into account;
- Arabic words can be made up of one or more Arabic word parts (PAW). A linked component known as a PAW can be a diacritical mark, a single letter, a string of letters, or an entire word;
- Arabic letters can be ligatured either vertically or horizontally. As a result, segmentation is a challenging process.

Challenges due to historical documents:

- Degradations in chemical composition brought on by changes in temperature, humidity, light, and air pollution. The paper may get yellowed as a result of these chemical reactions, or inks and pigments may cause the paper to become discolored;
- Human degradations brought on by people, such as notes added to documents and scratches, etc., or even the existence of a prior script, totally or partially erased (palimpsest);
- Degradation induced by living beings, such as rats or insects;
- Deterioration brought on by the digitizing process, including changes to resolution, support, compression standards, etc;

All the aforementioned challenges make it difficult to set a standard way to process historic documents, in particular Arabic-character ones.

Datasets

Despite the extensive scale of the historical document processing task, datasets for training, testing, and evaluation remain scarce in comparison to those used in related fields such as modern handwriting recognition. For Western historical documents, research datasets exist for medieval Latin, medieval German and Spanish, a variety of early modern European languages, and eighteenth-century English.

However, for Arabic, only few historic Arabic document datasets exist in the literature. In the following, a summary of the most relevant ones, as reported by (Khedher 2020).

WAHD dataset. It is a dataset (Abdelhaleem 2017) that was made public in 2017 at the University of the Negev. with the purpose of testing writer categorization. It is made up of 353 individual manuscripts that were gathered from two sources: the National Library in Jerusalem (NLJ; 333 manuscripts) and the Islamic Heritage Project (IHP; 20 manuscripts).

VML-HD dataset. It is a dataset (Kassis 2017) for word finding and word recognition tasks that was released in 2017 at the University of the Negev. The dataset consists of five manuscripts, each of which was produced between the years 1088 and 1451 by a different scribe. The 680 pages of the dataset are completely annotated at the sub-word level. There are 121,636 sub-word occurrences in all, spread across 1,731 classes.

HADARA80P dataset. It was proposed in 2014 by Pantke (2014) at TU Braunschweig. It was retrieved from a historical book untitled “Taaun” that contains 80 pages in total. In addition to scanned images, XML files with the ground truth are made available. 16,720 annotated words are present in the dataset. Additionally, it offers a choice of 25 pre-defined keywords. Between 5 and 349 occurrences of the complete match keywords may be found in the dataset.

IBN-SINA dataset. It was proposed by Farrahi (2010) at ETS Montreal, based on historic documents given by the Institute of Islamic Studies (IIS). To extract 20,722 forms, the dataset was manually annotated at the sub-word level (connected components). The shapes may represent letters, words, or subwords. The dataset may have more than 1,000 basis linked components based on clustering methods (keywords). The dataset is useful for word recognition and word location applications.

BADAM dataset. 42 manuscripts from four digitized collections of the Arabic and Persian languages are included in the public BADAM dataset (Kiessling 2019). 400 annotated pages from various disciplines and eras are included. Each manuscript had 10 pages selected from it, with the exception of four shorter manuscripts that only had 3-7 pages.

MHDID dataset. It stands for Historical Document Image Database with Multiple Distortions, and contains 335 historical document images with a size of 1024×1280 pixels (Shahkolaei 2018). It has a variety of distortions that are loosely divided into four categories: i) Paper translucency (88 photos), ii) Stain (113 images), iii) Readers comments (61 images), and iv) Worn holes (73 images).

Archive of the University of Napoli L'Orientale. It consists in a large collection of historic manuscripts in Arabic character, for several oriental languages, such as Arabic, Farsi, ottoman Turkish, etc. that has been digitized, including vocalized and unvocalized texts, which a specific and very relevant issue in Arabic-character recognition. We believe that such a material may be useful to start developing a standard dataset, and also to develop techniques for Arabic-character HDP. An example of this archive is shown in Fig. 1.

Fig. 1 - A scanned copy of a page from a Koran manuscript (Kor. 2:54-74). This picture illustrates some of the challenges in Arabic-character HDP: Arabic character is cursive; black dots are due to ink stains, red characters do not belong to the text, but indicate how to perform Koranic recitation, courtesy of SiBA-UniOr

Methods

The workflow of HDP involves several procedures or phases for digitizing the pages of a manuscript or an early printed book. After the document has been scanned, the page scans are generally preprocessed as follows: i) Binarization and thresholding for grayscale images, ii) analysis and segmentation of the layout, and iii) Text line normalization.

Then, utilizing machine learning software, optical character recognition (OCR) or handwritten text recognition (HTR) is performed, depending on the kind of document, as illustrated in the following figure (Fig. 2).

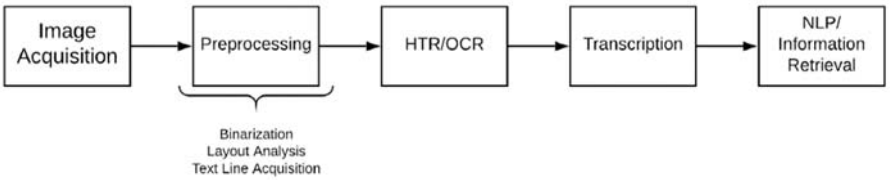

Fig. 2 - Conventional Historical Document Processing workflow (Philips 2020)

Prior to segmentation and character recognition, pre-processing is applied to try to normalize the documents, increasing the chance that the segmentation will be successful. The segmentation techniques employed determine the approach to be used; one popular pre-processing technique converts the image to black and white. Binarization (O'Gorman 1994) facilitates the separation of the document's components and enhances the text's definition and curvature (Fig. 3). To change the direction of text, techniques like skew angle correction (Al-Khatatneh 2015) are employed. The Hough Transform (Hough 1962) is one of the alternatives that uses parametric line representations to identify the lines in an image.

Fig. 3 – Unbinarized vs. binarized manuscript images (Jana 2017)

Principal Component Analysis (PCA) is used with binary images. The black pixels in an image are represented by a 2-dimensional vector. A projection profile that can be derived from this format may be used to represent a collection of 2-dimensional black pixel vectors. These vectors will point in the rotational axis of the picture. In light of this, the difference in skew angle is determined (Basavanna 2015; Fig. 4).

Filtering: Median filters and denoising filters are used to repair document degradation. Unwanted degradation occasionally happens as a result of improper image processing and scanning. This degradation or noise may have a damaging effect on segmentation algorithms. Simple filters like median and gaussian blur are widely used to denoise document image files. An image with a Gaussian blur may be created by simply convolution with a 2-dimensional vector whose pixel values sum to a Gaussian distribution. This process produces a smoothing effect and softens the borders and other aspects of a picture (Kumar 2013). Alternatively, median filters are used to replace each pixel in a picture with the median value of the pixels around it. This is done by creating a window through the image. The sliding window's size can be changed according to how much "smoothing" is necessary to fix the image.

A scanned document image is divided into its smallest components by a technique called segmentation. Each component in a portion comes from the block before it. Prior to character

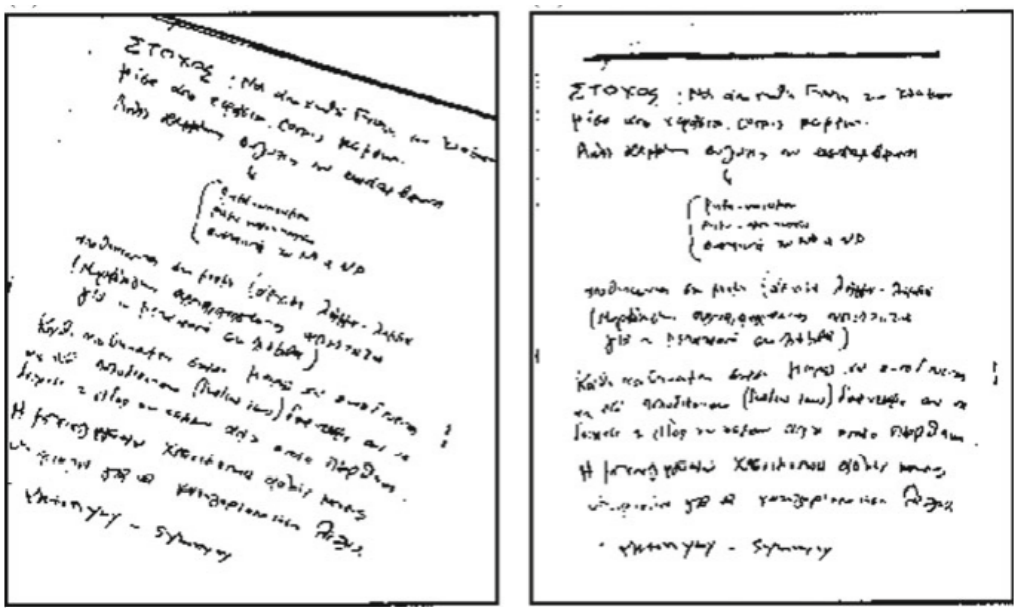

Fig. 4 - Incorrectly vs. correctly oriented text (Bugeja 2020)

recognition, a document image is first segmented into lines, words, and characters (document segmentation). This method is also utilized in the industrial document digitization process, which uses similar transcription techniques as well as additional document image analysis methods including keyword spotting. Techniques for document segmentation algorithms go beyond just partitioning text into lines, words, and characters. To separate and extract contexts like images, headlines, and articles, certain approaches are utilized.

Various elements that make up document images employ various methodologies. An overview of the methods utilized at each stage of the segmentation process is covered in this section. All stages of segmentation share a few of the same strategies.

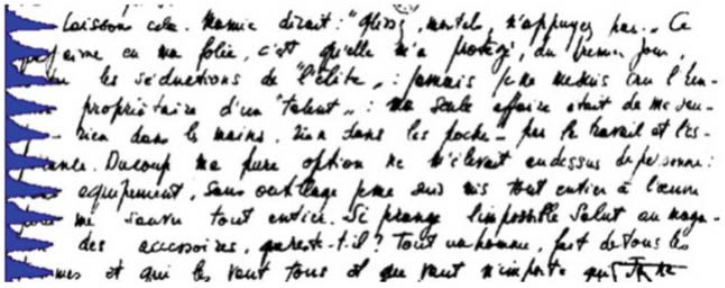

Fig. 5 - Projection profile segmentation (Bugeja 2020)

The projection profile technique is one of these strategies. These methods extract white space areas from a document picture using histograms. These techniques demand that the document pictures be appropriately oriented and translated to a binary form. The text lines are divided into segments by the white lines. Using projection profiles presents a challenge since lines cannot always be divided by a straight line in handwritten text (Ouwayed 2012). Additionally, text lines in handwritten text may overlap. Thus, it is impossible to compute a perfect straight line that can accurately divide text into its component words, lines, and characters. With varying degrees of effectiveness, the same procedure is also applied to character segmentation and word segmentation (Louloudis 2009; Fig. 5).

White space analysis approaches, in contrast to projection profile techniques, compute the space between lines and words in a document picture by generating a histogram that counts the number of

white pixels in a binarized image. The amount of minimal space between linked components is calculated using white space analysis. These linked elements are then arranged vertically to form words or horizontally to depict lines into chains. In order to calculate white lines and column dividers, the related components are grouped vertically as a result of this method (Chen 2013).

Fig. 6 - Application of erosion and dilation techniques (a, b, c, d, e) followed by Gaussian blur (f) to extract lines (Bugeja 2020)

Smoothing techniques, such as Gaussian blur, can enhance line extraction by removing noise caused by erosion and dilation (Bockholt n.d.; Boiangiu 2013; Fig. 6). Most Hybrid segmentation approaches employ filters. The Steerable Directional Local Profile Technique is an algorithm that was nonetheless created by the authors of Shi (2009). This method is based on their observations of how people extract lines manually. By seeing text line patterns, humans frequently extract lines from document images. To modify the size of a document picture, an adaptive local connectivity map (ALCM) feature is employed. Each pixel's horizontal directions are calculated using the intensities obtained by interconnecting its neighbors' pixels. A local adaptive thresholding approach is utilized to identify the text line patterns present in the collected ALCM with respect to their connected component.

Edge detection is a method used in computer vision to locate an object's or shape's edge within an image. These methods have also been effectively used to identify the edges of words, lines, and characters in handwritten text. Finding the boundary between these parts can be challenging, especially in manuscripts with overlapping lines, words, and recursive language.

Clustering techniques determine the distance between joined parts in both the horizontal and vertical planes. The method then makes the assumption that short distances on the vertical space between linked components are on the same line. The next or prior text line is supposed to be the distance between linked components on the horizontal line.

Supervised learning algorithms are frequently utilized to recognize related components that overlap and to segment them appropriately. Then, each component is combined into a graph component. A graph's nodes are connected by edges, and the weight of an edge depends on how far apart the edges are from one another. The lines are then removed from the graph by looking for the line that is stated as being the straightest (Liu 2014). Top-down methods called "function analysis" aim to estimate the objective value of a function (a segmented line or word). Functions including probabilistic layout

estimation, contours, and energy mapping are computed. Instead of border regions, these algorithms often operate on boundary edges.

Once segmentation is achieved, character recognition can be performed based on AI classification methods. These models rely on a collection of appropriately segmented character images as their input. However, the images are reduced to a set of characteristic features.

The feature set used to train the model is crucial to establishing the overall classification accuracy in machine learning. The model's accuracy increases with the quantity of features, the size of the dataset, and the quality of the features. Additionally, the model becomes more generic the more diverse and expansive the dataset is.

Feature selection and normalization is a necessary step to select standard feature sets utilized in handwriting recognition, e.g. diagonal features, contour features, and geometrical features. By splitting the image into bins and counting the number of diagonals at each bin, diagonal characteristics are recovered from the image. Utilizing counter detection methods like Freeman Chain Codes, counter characteristics are retrieved (FCC). These algorithms take a character image's edge direction into account. The amount of characteristics retrieved is then normalized using a normalizing technique.

Latent feature extraction can be more useful in some cases than standard feature usage. Such features can be obtained from a binary image using K-means clustering. K-means clustering has the benefit of allowing for greater robustness in poor light. Additionally, because the feature set's dimensions are relatively small, there is little computational overhead. K-means clustering is used for each binary image in the data set.

Despite the distinctions between ancient manuscripts and printed works, text extraction problems may be solved in a similar way using optical character recognition (OCR) and handwriting recognition. After preprocessing and segmentation, text recognition either extracts keywords from a line of text or converts it into a verbatim transcription.

The main goal is to correctly convert the words in the document image into digital text, whether the content was written or printed. Optical character recognition bases its categorization on the predicted regularity of the space between letters and words. The essential unit of recognition is the character. Optical character recognition classifiers can identify and create a transcription using the individual character glyphs because words and their constituent characters can be segmented regularly and precisely. However, because of the peculiarities of human handwriting, handwriting recognition cannot rely on consistent letter and word spacing.

Sayre's paradox states that while segmentation is necessary for individual letter identification, previous recognition is also necessary for segmentation (Fischer 2011). Therefore, handwriting identification relies on a recognition process that is segmentation free at the character level (in contrast to optical character recognition).

If the document language is known in advance, systems will frequently use a statistical language model to increase the accuracy of both optical character recognition and handwritten text recognition (Frinken 2013). A language model ensures that the words the software recognizes match the language's syntax and even its known vocabulary. Both conventional machine learning methods and deep learning methods based on neural networks may be used for handwriting recognition and optical character recognition.

Historical document processing involves approaches for confirming the accuracy and effectiveness of the algorithms and software tools utilized, much like any data-driven digital investigation. Because historical interpretation depends on the quality of authenticity, data integrity is crucial in this subject. The majority of research and tools use either traditional machine learning

approaches or neural network-based techniques for handwriting recognition and optical character recognition. This implies that in order for the character or word classifier to map the text in the document image to its transcription during the training phase, it needs annotated transcription data, often known as “ground truth” (Philips 2020).

The majority of the technologies now in use employ supervised machine learning, which depends on annotated data for training and assessment. The Ocular OCR engine from the Natural Language Processing Lab at the University of California, Berkeley is an exception (Berg-Kirkpatrick 2013). This computer program makes use of an unsupervised classifier. Nevertheless, the annotated data is still required for the testing stage in order to evaluate the tool's performance and correctness on omitted data. Some research works divide the dataset into training, testing, and development subsets according to the traditional tripartite partitioning method. Some people use cross-fold validation methods. Some datasets, including the IMPACT and Diva-HisDB databases, also offer ground truth for layout analysis. A feature in OCRopus allows users to enter a ground truth transcription for text lines that have been retrieved from an HTML page.

A historical document processing system's performance is evaluated using a variety of metrics. Precision and recall are two crucial performance metrics for image-based handwritten text recognition systems. How many of the dataset's relevant results were really retrieved depends on precision. Character error rate, word error rate, or occasionally both are used by machine learning systems to assess the success of transcription when a language model is used to improve recognition outcomes. Line error rate and segmentation error rate are used to evaluate the effectiveness of layout analysis (Bosch 2014).

Potential applications

A software tool for Arabic-character HDP may go beyond manuscript transcription, to reach a wide variety of interesting patterns, categorization, and analytical issues that may be found while processing historical Arabic documents. Once the character recognition procedure is completed successfully, there are several ways to enhance its use by research, industrial and cultural communities, since it may be the basis to achieve other tasks, as reported by (Khedher 2020):

A historical document can’t be utilized directly since its quality is frequently reduced. To separate the document's core content from the remainder, layout analysis and line segmentation are necessary. This technique, which we refer to as Document Analysis, is a pre-processing phase that comes before writer categorization or data retrieval.

A historical document frequently omits several pages that contain crucial data, such the document’s metadata (year, writer, title, etc.). When the author is unidentified, we might try to find out more about him. To start, Writer Identification (WI) may be used to compare the work to other library documents with the same author. Second, one can get documents images created by the same writer - even an unknown one - from the database (Writer Retrieval (WR)). These two processes are referred to as Writer Classification.

It is sometimes challenging to identify each figure individually in the context of historical sources. However, by taking into account a higher degree of granularity - a sub-word, a word, or even the entire text, it is feasible to draw certain conclusions about the document, in order to achieve i) word recognition, ii) text alignment, i.e. the comparison of the transcript to other sources, and iii) data retrieval, such as searching a key word.

Scholars are able to identify language for historic documents. However, in certain cases, Arabic-character manuscripts may contain passages or side comments with another language, e.g. Arabic in a

Farsi documents. Therefore, a process of language identification may be performed on the recognized texts to detect any part written in a different language.

Once the historic document is converted into plain text, it becomes easy to convert it into sound using TTS techniques (Text-To-Speech) or to translate it to a modern language. Such applications would be very useful in museums or online galleries to make historic documents more comprehensible by the public, thus to increase its cultural impact.

Conclusion

In this paper, an overview of Arabic-character HDP has been presented. The main goal is to introduce this topic from different sides, including datasets, challenges, methods, and potential applications. It may particularly be interesting to investigate such aspects when a project aiming to achieve this task is to be launched. Through this literature review, the following remarks can be drawn: a) Datasets of scanned historical Arabic-character manuscripts are not numerous, however they offer a rich and a varied collection of input data; b) Challenges are mainly related to the inherent difficulty to segment the Arabic-character manuscript text, and also to the quality of the manuscripts; c) Methods are mainly related to image processing and pattern recognition techniques. However, at this end, machine learning is a good and a reliable tool; d) Potential applications are numerous, and span a wide range from document analysis, writer recognition to automatic translation and speech synthesis. Finally, we think that addressing the issues and challenges mentioned may help improving Arabic-character HDP technology, contributing in the dissemination of its cultural heritage.

REFERENCES

Abdelhaleem, A., A. Droby, A. Asi, M. Kassis, R. Al Asam, J. El-sanaa (2017) Wahd: a database for writer identification of arabic historical documents, *1st International workshop on arabic script analysis and recognition* (ASAR), pp 64-67, IEEE.

Al-Khatatneh, A., S. A. Pitchay, M. Al-qudah, (2015) A review of skew detection techniques for document, *17th UKSim-AMSS International Conference on Modelling and Simulation*, pp 316-321, IEEE. ("A Review of Skew Detection Techniques for Document").

Asi, A., R. Cohen, K. Kedem, J. El-Sana, I. Dinstein, (2014) A coarse-to-fine approach for layout analysis of ancient manuscripts, *14th International Conference on Frontiers in Handwriting Recognition*, pp 140-145, IEEE.

Basavanna, M., S. S. Gornale (2015) Skew detection and skew correction in scanned document image using principal component analysis, *Int. J. Sci. Eng. Res, 6(1)*, pp 1414-1417.

Berg-Kirkpatrick, T., G. Durrett, D. Klein (2013) Unsupervised transcription of historical documents, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistic,*. pp 207-217.

Biller, O., A. Asi, K. Kedem, J. El-Sana, I. Dinstein(2013) Webgt: An interactive web-based system for historical document ground truth generation, *12th International Conference on Document Analysis and Recognition*, pp 305-308. ("Ontology-based semantic search development on Lanna King History using ...").

Bockholt, T. C., G. D. Cavalcanti, C. A. Mello(n.d) Document image retrieval with morphology-based segmentation and features combination, *Document Recognition and Retrieval XVIII 7874*, pp 356-367.

Boiangiu, C. A., M. C. Tanase, R. Ioanitescu (2013) Text line segmentation in handwritten documents based on dynamic weights, *Journal of Information Systems & Operations Management (JISOM)* vol.7. n.2, pp 247-254.

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

Boussellaa, W., A. Zahour, B. Taconet, A. Alimi, A. Benabdelhafid PRAAD (2007) Preprocessing and analysis tool for Arabic ancient documents, *Ninth International Conference on Document Analysis and Recognition (ICDAR)*.

Boussellaa, W., A. Zahour, B. Taconet, A. Benabdelhafid, A. Alimi (2006) Segmentation texte/graphique: Application au manuscrits Arabes Anciens, *Actes du 9ème Colloque International Francophone sur l'Ecrit et le Document*, pp 139-144.

Bugeja, M., A. Dingli, Seychell (2020) *An Overview of Handwritten Character Recognition Systems for Historical Documents*.

Chen, K., F. Yin, C. L. Liu (2013) Hybrid page segmentation with efficient whitespace rectangles extraction and grouping, *12th International Conference on Document Analysis and Recognition*, pp 958-962, IEEE.

Cohen, R., I. Rabaev, J. El-Sana, K. Kedem, I. Dinstein (2015) Aligning transcript of historical documents using energy minimization, *13th International Conference on Document Analysis and Recognition (ICDAR)*, pp 266-270, IEEE.

Farrahi M., M. Cheriet, M. M. Adankon, K. Filonenko, R. Wisnovsky (2010) IBN SINA: a database for research on processing and understanding of Arabic manuscripts images, *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pp 11-18.

Fischer, A., E. Indermuhle, V. Frinken, H. Bunke (2011) HMM-based alignment of inaccurate transcriptions for historical documents, *International Conference on Document Analysis and Recognition (ICDAR)*, pp 53-57, IEEE.

Frinken, V., A. Fischer, C. D. Martínez-Hinarejos (2013) Handwriting recognition in historical documents using very large vocabularies, *Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing*, pp 67-72.

Hough, P. V. (1962) The Hough transform, *U.S. Patent No. 3,069,654.* Washington, DC.

Jana, P., S. Ghosh, S. K. Bera, R. Sarkar (2017) Handwritten document image binarization: An adaptive K-means based approach, *IEEE Calcutta Conference (CALCON)*, pp 226-230, IEEE.

Kahle, P., S. Colutto, G. Hackl, G. Mühlberger (2017) Transkribus-a service platform for transcription, recognition and retrieval of historical documents, *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, IEEE. ("Improving Handwriting Recognition for Historical Documents Using ...").

Kassis, M., A. Abdalhaleem, A. Droby, R. Alaasam, J. El-Sana, (2017) Vml-hd: The historical arabic documents dataset for recognition systems, *1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, pp 11-14, IEEE.

Khader, H., A. Al-Marridi, H. Alpona, S. Kunhot, A. Hassaine, S. Al-Maadeed (2014) An interactive annotation tool for indexing historical manuscripts, *World Symposium on Computer Applications & Research (WSCAR)*, pp 1-4, IEEE.

Khedher, M. I., H. Jmila, M. A. El-Yacoubi (2020) Automatic processing of Historical Arabic Documents: a comprehensive survey, *Pattern Recognition* vol.100, pp 107-144.

Kiessling, B., D. S. B. Ezra & M. T. Miller (2019) BADAM: a public dataset for baseline detection in Arabic-script manuscripts, *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*, pp 13-18.

Kumar, B. K. (2013) Image denoising based on non-local means filter and its method noise thresholding, *Signal, image and video processing* vol.7, n. 6, pp 1211-1227.

Liu, L., Y. Lu, C. Y. Suen (2014) Near-duplicate document image matching: A graphical perspective, *Pattern recognition* vol.47, n. 4, pp 1653-1663.

Louloudis G., B. Gato, I. Pratikaki, C. Halatsi (2009) Text line and word segmentation of handwritten documents, *Pattern Recognition* vol.42, n.12, pp 3169–3183.

O'Gorman, L. (1994) Binarization and multithresholding of document images using connectivity, *Graphical models and image processing*, *vol 56, n.6,* pp 494-506.

Otsu, N. (1979) A threshold selection method from gray-level histograms, *IEEE transactions on systems, man, and cybernetics* vol.9, n.1, pp 62-66.

Ouwayed, N., A. Belaïd (2012) A general approach for multi-oriented text line extraction of handwritten documents, *International Journal on Document Analysis and Recognition (IJDAR)*, vol.15, n.4, pp 297-314.

Pantke, W., M. Dennhardt, D. Fecker, V. Märgner, T. Fingscheidt (2014) An historical handwritten Arabic dataset for segmentation-free word spotting-HADARA80P, *14th International Conference on Frontiers in Handwriting Recognition*.

Philips, J., N. Tabrizi (2020) Historical Document Processing: A Survey of Techniques, Tools, and Trends, *KDIR* pp 341-349.

Povey, D., A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, K. Vesely (2011), The Kaldi speech recognition toolkit, *IEEE workshop on automatic speech recognition and understanding*, 2011, IEEE Signal Processing Society. ("Sichuan dialect speech recognition with deep LSTM network").

READ Project. <https://readcoop.eu/transkribus/>.

Shahkolaei, A., A. Beghdadi, S. Al-Máadeed, M. Cheriet (2018) Mhdid: a multi-distortion historical document image database, *2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, pp 156-160, IEEE.

Shi, Z., S. Setlur, V. Govindaraju (2009) A steerable directional local profile technique for extraction of handwritten arabic text lines, *10th International Conference on Document Analysis and Recognition*, pp 176-180, IEEE.

SiBA-Unior, Sistema Bibliotecario di Ateneo, University of Naples L'Orientale, n.d.

Srihari, S. N., V. Govindaraju (1989) Analysis of textual images using the Hough transform, *Machine vision and Applications* vol.2, n.3, pp 141-153.

Stahlberg, F. & S. Vogel (2016) QATIP-An Optical Character Recognition System for Arabic Heritage Collections in Libraries, *12th IAPR Workshop on Document Analysis Systems (DAS)*, pp 168-173, IEEE.

Würsch, M., R. Ingold, M. Liwicki (2017) DIVAServices - A RESTful web service for Document Image Analysis methods, *Digital Scholarship in the Humanities* vol.32, n.1, pp 150-156.

Zahour, A., B. Taconet & S. Ramdane (2004) Contribution à la segmentation de textes manuscrits anciens, *Conférence Internationale Francophone sur l'Ecrit et le Document (CIFED).*