A survey on text-line segmentation process in historical Arab manuscripts

Soumia Djaghbellou[1],*, Attia Abdelouahab[2] and Bouziane Abderraouf[3]

[12] *Department of computer science, university of Bordj Bou Arreridj, Algeria*
[3] *MSE Laboratory, Department of computer science, university of Bordj Bou Arreridj, Algeria*

Abstract

The segmentation process entails dividing or decomposing the entire document image into segments or lines. This technique serves as a fundamental step in developing any writing or optical character recognition system. However, numerous existing segmentation schemes encounter challenges when dealing with specific script styles, like ancient or historical Arabic writing found in ancient manuscripts. which possesses unique characteristics. These characteristics include inclined text lines, overlapping letters, diacritic marks, decorative elements, variable letter forms, and ligatures (combinations of two or more letters merged to form a single connected shape).Thus, in this paper we present a thorough survey of the field. The survey is composed of two segments. The first segment provides a concise overview of the historical Arabic documents. The second, which serves as the primary segment, focuses on the crucial step of handwritten document recognition, specifically segmentation. A detailed and systematic overview of the various approaches to segmentation, including different levels, employed for extracting handwritten Arabic text-lines, is outlined. Subsequently, a literature study is conducted to review and analyze proposed works in this area.

Keywords

Text-lines, segmentation, pattern recognition, Arabic handwritten, historical Arabic documents

1. Introduction

Historical documents often symbolize the identity of diverse civilizations worldwide. Analyzing and comprehending their contents holds paramount importance, especially for researchers. Manually extracting data from these historical documents proves to be a laborious and expensive endeavor. In recent years, there has been a surge in research dedicated to the automated processing of historical documents. Despite notable advancements, automating the processing and analysis of historical Arabic documents remains a challenging task. Text line segmentation stands as an initial stage in the text recognition system process. This critical preprocessing step in document analysis poses particular challenges, especially with handwritten texts. While segmenting text lines from machine-printed documents is commonly considered resolved, freestyle handwritten text lines remain notably challenging. This complexity arises due to curved lines, inconsistent spacing, and overlapping spatial boundaries. Additionally, irregular layouts, varied character sizes reflecting different writing styles, intersecting lines, and the absence of a clear baseline all contribute to the intricacy and difficulty in handwritten document analysis[1]. This paper focuses on complex problem of text-line segmentation process and provides a comprehensive survey of existing research works. Firstly, it presents the historical Arabic Manuscripts in general, including main features, structure/type of documents. Secondly, in the main section of the paper, the focus shifts to the segmentation process as a critical phase in recognition, specifically emphasizing the commonly utilized and adopted techniques for Arabic scripts. The remainder of this paper is structured as follows: section 2 offers a comprehensive overview of ancient Arabic manuscripts, encompassing their content structure and various applications. Moving on to Section 3, we delve into the image segmentation process, exploring its different levels and focusing on widely adopted approaches specifically designed for handwritten Arabic texts. In Section 4, we concentrate on a comparative study, presenting notable existing works related to the segmentation of handwritten Arabic documents. This analysis will consider the method of experimental analysis and the data-set used. Section 5 is dedicated to showcasing a compilation of famous Arabic databases that have been utilized in various studies. Finally, Section 6 addresses open issues, motivations, and potential directions for future research. And lastly, Section 7 concludes the paper, summarizing the findings and insights presented throughout the article.

6th International Hybrid Conference On Informatics And Applied Mathematics, December 6-7, 2023 Guelma, Algeria

*Corresponding author.
soumia.djaghbellou_1@univ-bba.dz (S. Djaghbellou); attiaabdelouahab@gmail.com (A. Abdelouahab); bouzianeabderraouf@gmail.com (B. Abderraouf)

2. Structures and applications of the historical Arabic documents

Throughout history, manuscripts were the official way of writing down knowledge and science. There is a huge amount of historical Arabic manuscripts in the archives and national libraries around the world, which have been scanning their collections to make them publicly available and to preserve this valuable cultural heritage. The Arabic manuscripts, like manuscripts of the other languages, have some common characteristics.

Manuscripts have their specificities and various distinct elements that can useful to identify the manuscripts for the creation of an electronic format of description. Some of these elements, we distinguish are:

- Mention the responsible
- Names of owners. It is also an important clue for researchers wishing to follow the historical development of the manuscript
- The title: it is the main identifying element which in most cases is presented on the title page
- Physical description/codicology.

Handwritten documents, irrespective of the language, are typically categorized based on their physical appearance into four classes, as outlined in [2]:

- Mono-oriented documents: lines in this class are oriented in one direction. Figure 1.a shows a handwritten Arabic document with a horizontal orientation
- Multi-oriented documents: lines in these documents are arranged in blocks of different orientations. Figure 1.b gives an example of this class of documents
- Multi-script documents: These comprise texts authored by multiple individuals, resulting in various scripts. This occurrence was frequent in the past when individuals succeeded one another to finalize a document or collaborated on the same written piece. Figure 1c depicts a multi-script handwritten document (Arabic and Latin)
- Heterogeneous documents: This category encompasses content that includes both textual information and images or illustrations. Handwritten documents of this nature might feature diverse orientations, such as dimension lines or illustrative drawings, as illustrated in Figure 1d.

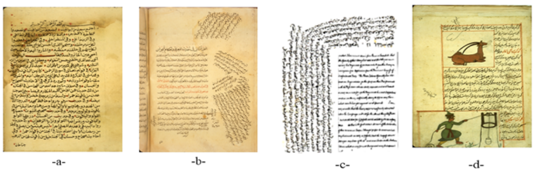


Figure 1: Examples of the four categories a,b,d [3], d [4] of handwritten documents.

Currently, numerous major libraries globally are digitizing handwritten historical documents. These scanned images are uploaded onto their websites, complemented by metadata facilitating specific document searches within extensive databases. Accessing document content isn't feasible without its digital presence in a textual format. Therefore, to harness these documents, diverse methods and applications are employed. In this paper, we succinctly outline three applications, depicted in the figure 2 [5]. In dealing with aged document images, segmentation poses a significant challenge. However, delineating distinct blocks within their physical structure simplifies this task. By focusing on block forms and their spatial relationships, we ascertain:

- Baseline (connecting the lower part of character bodies)
- Median line (tracing the upper part of character bodies)
- Upper line (linking the top of ascenders)
- Lower line (uniting the bottom of descenders).

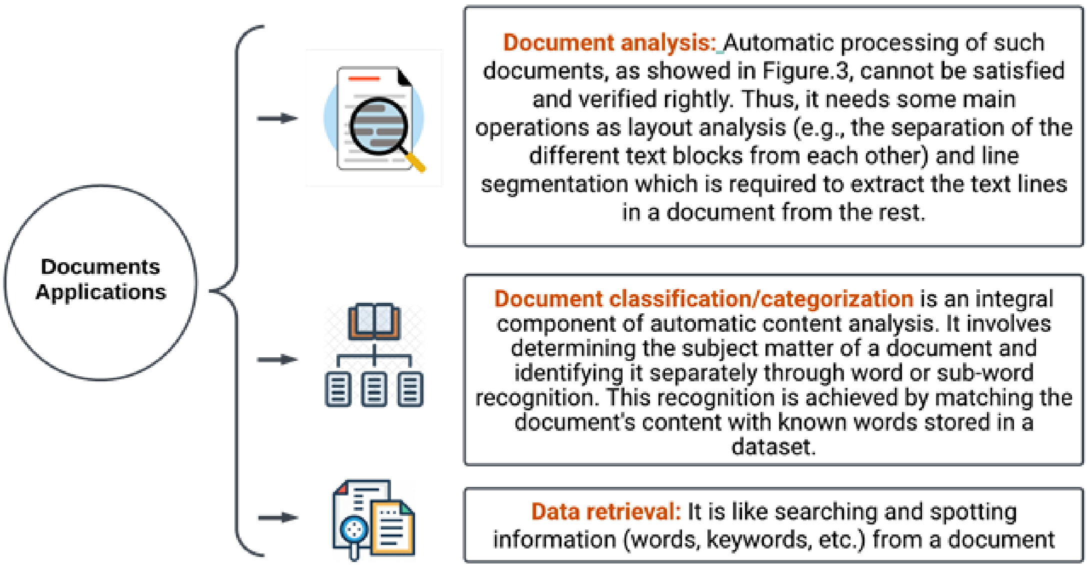
Effectively interpreting this manuscript type demands a robust segmentation process backed by efficient methods and techniques. This paper aims to extensively discuss the pivotal phase of segmenting textual documents, concentrating on the most effective methodologies that have exhibited performance, particularly concerning handwritten Arabic script.

3. Segmentation phase and its methods

Segmentation of documents into text lines, also known as text line extraction, stands as a fundamental step in document content recognition. It typically serves as a preprocessing phase, as illustrated in Figure 4.

However, segmenting ancient and historical handwritten Arabic documents into text lines poses considerable challenges, especially when dealing with documents of poor quality. This complexity hampers content extraction due to the diverse nature of Arabic writing. Characters and words present varying shapes, contributing to an extensive vocabulary. Moreover, these documents frequently feature additional disruptive elements like stains, ornamentation, seals, and holes [5].

The objective of segmentation is to simplify and alter the representation of an image into something more meaningful and easier to analyze and recognize, such



Document analysis: Automatic processing of such documents, as showed in Figure.3, cannot be satisfied and verified rightly. Thus, it needs some main operations as layout analysis (e.g., the separation of the different text blocks from each other) and line segmentation which is required to extract the text lines in a document from the rest.

Document classification/categorization is an integral component of automatic content analysis. It involves determining the subject matter of a document and identifying it separately through word or sub-word recognition. This recognition is achieved by matching the document's content with known words stored in a dataset.

Data retrieval: It is like searching and spotting information (words, keywords, etc.) from a document

Figure 2: The main documents applications.

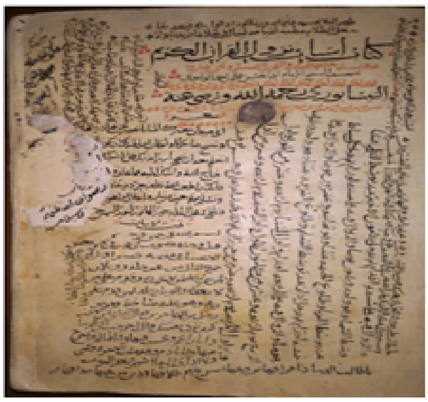


Figure 3: Arabic document of different text blocks[4].

as lines, sub-words, words, or characters. This is illustrated in Figure 5. In general, there exist four levels of segmentation, as illustrated in Figure 5:

Page segmentation: This initial step involves identifying information areas on each page based on their visual attributes. It often includes logically labeling these areas according to the content they represent, such as text, graphics, or images. A comprehensive analysis of the techniques employed in document analysis has been presented in prior studies [6] [7] [8].

Text segmentation into lines: This stage focuses on separating text lines to extract individual words and subsequently, the characters within those words. Numerous studies in this domain employ image decomposition into connected components [9].

Line segmentation into words: This phase utilizes vertical projections' histograms of lines to detect spaces between words for separation. However, this method might not be as effective when dealing with Arabic script.

Word segmentation into characters: This process involves breaking down words into their constituent individual symbols. It is a pivotal decision-making step in optical character recognition systems, determining the accuracy of isolated patterns within an image [10].

3.1. The adopted methods for Arabic text-lines segmentation

In the context of Arabic manuscripts, localizing text lines for extraction or segmentation is challenging due to the morphological peculiarities associated with the Arabic script. The script is naturally cursive, unconstrained, and horizontally oriented, which adds to the difficulties, especially when dealing with historical documents. Evaluation of Arabic handwritten text line extraction algorithms has either been lacking or has shown higher error rates compared to algorithms used for other languages. This is because, as previously mentioned, Arabic script exhibits greater cursive characteristics compared to other scripts [11].

To streamline and simplify the handling of such documents, extensive research has been directed toward

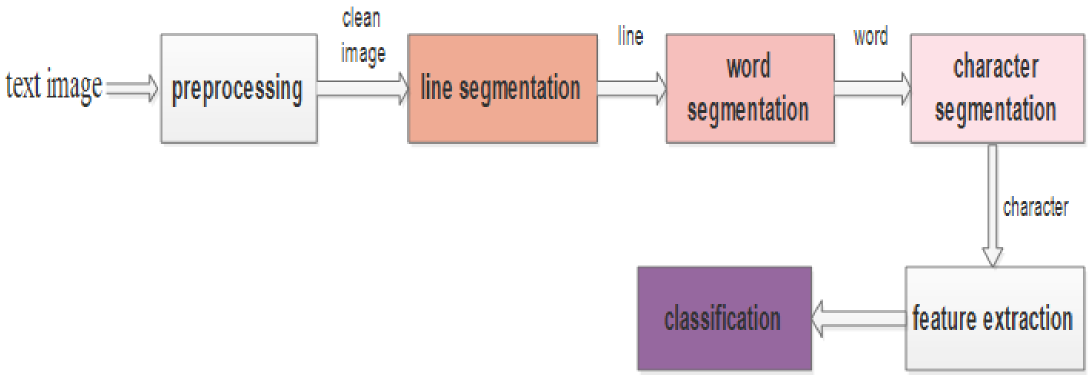


Figure 4: Text Recognition Process

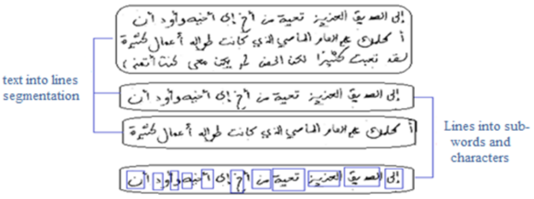

text into lines segmentation
Lines into sub-words and characters

Figure 5: The text segmentation levels [10].

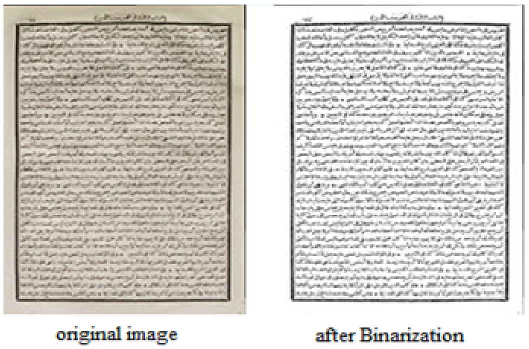

original image
after Binarization

Figure 6: Original colored document and Binarized document.

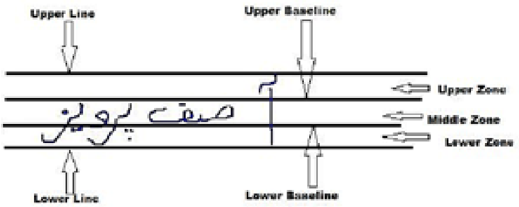


Figure 7: Reference lines and interfering lines [17].

image binarization. Image binarization, illustrated in Figure 6, involves dividing the image into two categories: the background and the object (text lines). Segmenting a color image can be exceptionally resource-intensive, often necessitating consideration of multiple factors (e.g., histogram, color, etc.) to determine a point's class or type. Various techniques have been devised for image binarization, including global threshold, local threshold, Markov random field model [12], [13], water flow model [14], and Gatos et al.'s method [15], [16]. In general, methods and approaches for segmenting handwritten text lines can be categorized based on their operational mode or the strategies they employ. These strategies cover projection-based methods, smearing methods, hough-based approaches, and clustering or grouping methods.

Projection-based Approach: Within this approach, we differentiate between two profiles: the vertical projection and the horizontal projection. These profiles are obtained by summing the pixel values along the vertical and horizontal axes, respectively, for each y and x value [15]. In our text line segmentation study, we specifically focus on the horizontal projection. This projection is applied to different lines in order to obtain an initial position for the separated lines and their corresponding baselines. In Figure 7, we present the reference lines and interfering lines, while Figure 8 provides an illustrative example of an Arabic text and its horizontal projection.

Smearing methods: Involve horizontally spreading black pixels while measuring the gap between white spaces. If the distance meets a predetermined threshold, those spaces get filled with black pixels. This process leads to the formation of interconnected black pixel shapes around the text lines [18], as depicted in Figure 9. However, this method posed a challenge of line overlap. Consequently, during smearing, two separate lines might merge, causing the segmentation of two lines into

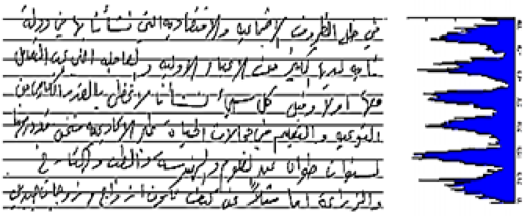


Figure 8: Arabic text and its horizontal projection [11]

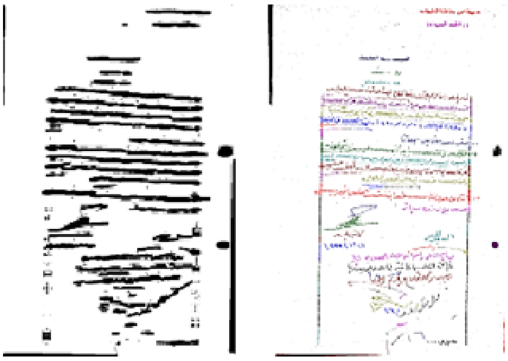


Figure 9: An example demonstrating the Application of the Smearing Method on Arabic text [18]

one[19].

Hough transform: is a frequency-based technique employed to detect and pinpoint straight lines within text document images. Employing the Hough transform on the centroid allows us to determine the orientations of these lines [20]. Handwritten document images often contain annotations, erasures, and lines oriented in various directions alongside the primary lines. Consequently, this method depends on contextual cues, like direction continuity and proximity criteria, to eliminate erroneous alignments among the components.

Grouping Approach: This method involves aggregating and combining various units (such as blocks, pixels, connected components) in a bottom-up fashion to create alignments using distinct perceptual criteria such as proximity, continuity, and similarity. It also employs geometric feature details, including the size, position, shape, and orientation of the connected components, thereby grouping them into rows [20].

Most studies on segmenting handwritten Arabic text into lines rely on schemes that identify overlapping, touched, or connected components, such as the grouping approach. However, projection-based methods remain effective primarily in cases where handwritten documents exhibit minimal overlap, and there is no whitespace between lines, a common feature in many Arabic manuscripts. Conversely, alternative methods (refer to Figure 10) are guided by criteria aimed at preventing the intersection of black pixels, as proposed and implemented in [21] for segmenting both modern and historical Chinese documents, often characterized by overlapping lines. Nevertheless, this method proves efficient only when fewer black pixels are present at the contact points, potentially failing when these points contain a significant amount of such pixels.

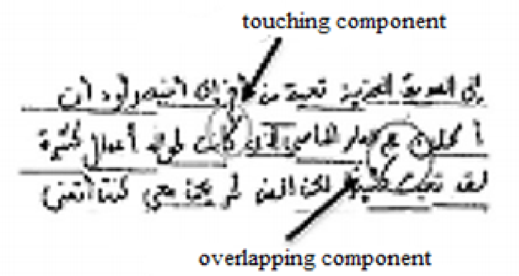


Figure 10: Overlapping and touching components [20]

For detecting connected components between lines, several rules and criteria need consideration. These include identifying the label of the component (touching/overlapping), managing ambiguous component sizes, assessing the density of black pixels within each alignment region, considering alignment proximity, and evaluating contextual information (such as the positions of alignments surrounding the component) [22]. Separating these components involves analyzing the vertical projection profile of the component to determine the location of the horizontal frontier segment that will be used to separate the touching elements. If the projection profile exhibits two peaks, the separation occurs midway between them; otherwise, the component is divided into two equal parts [22].

4. Literature on Handwritten Arabic Documents Segmentation works

In this section, we showcase prominent existing research in the realm of segmenting handwritten Arabic documents, employing experimental analysis methods and datasets. Until recently, there has been a scarcity of studies focusing on segmenting and recognizing handwritten Arabic text. Owing to the script's distinct characteristics, conventional methods often fall short in terms of efficacy. The review encompasses publications from the last 12 years, evaluating a total of 10 articles from journals and conferences. These articles primarily centered on text segmentation within their core methodologies. A summary of these ten articles is provided in Table 1. For a better understanding of the Table 1 contents, let's examine each column separately. The first column in the table displays the author and the publication year of the article. Next, a concise description of the segmentation technique outlined in each respective study is provided. Afterward, details regarding the database utilized to test the proposed model are presented. Lastly, the final column showcases the evaluation results of each experiment. It's important to highlight that the assessment of the performance of different proposed methods relied on calculating various metrics as detailed below:

Accuracy = (TP+TN)/ (TP+TN+FP+FN)

Precision = TP/ (TP+FP)

Recall=TP/ (TP+FN)

F1-Score=2*((Precision*Recall)/ (Precision+ Recall))

Where:

TP, TN (True Positives and True Negatives): indicate the correct predictions for the positive and negative class.

FP, FN (False Positives and False Negatives): FP indicates the incorrect predictions of the positive class and the incorrect predictions of the negative class, often referred to as FN.

Boussellaa, Wafa, et al. (2010) [23] introduced a segmentation method centered on block covering analysis employing an unsupervised method. Initially, they calculated the optimal document decomposition into vertical strips to achieve fuzzy baseline detection using the fuzzy C-means algorithm. Afterward, blocks were assigned to the respective lines in Arabic historical handwritten documents containing various scripts, including characters that overlapped or were multi-touching. The algorithm presented in their study demonstrated strong performance, achieving an accuracy rate of 95%.

Kumar, Jayant, et al. (2010) [24] introduced a method for extracting handwritten text lines from monochromatic Arabic document images. Their approach relied on a unique graph framework involving two key steps. Initially, the scheme estimates local orientation at each primary component, constructing a sparse similarity graph. Subsequently, it employs a shortest path algorithm to assess similarities between non-adjacent components. The model underwent testing on a dataset comprising 125 images, resulting in a final accuracy of 96% .

Kumar, Jayant, et al. (2011) [25] developed a handwritten text-lines extraction model utilizing a graph-based technique to detect touching and proximity errors, with a refinement step using Expectation-Maximization (EM) to iteratively split the error segments to obtain correct text-lines of the experiment dataset of 125 Arabic document images. The study showed the productivity of the proposed experiments by giving a very high score of 98.76% .

Khayyat, Muna, et al. (2012) [26] introduced a technique for extracting handwritten text lines. Their method relies on morphological dilation with a dynamically adaptive mask. They employed a smearing technique to generate large connected components, or blobs, which were subsequently analyzed for applying appropriate smearing to the document. This approach underwent testing using the Arabic dataset CENPARMI, encompassing multi-skewed and touching lines. The experimental outcomes demonstrated the efficacy of the proposed algorithm, achieving a precision rate of 96.3% .

Al-Dmour and Fares Fraij (2014) [27] introduced a text-line segmentation model that relies on the established horizontal projection profile (HPP) method. Initially, the approach involves generating a histogram of black pixels along the preprocessed image's horizontal scan lines. Subsequently, the self-similarity is improved through autocorrelation. The implemented system demonstrated highly encouraging outcomes, achieving an extraction accuracy rate of 84.8% .

Suresha, M., and Amani Ali Ahmed Ali (2018) [28] developed a segmentation process utilizing the Hough transform method. This technique was followed by a skeletonization operation in the post-processing phase, aimed at rectifying potential false alarms. The ultimate objective was to proficiently segment vertically connected characters. The effectiveness of the proposed system was demonstrated through experimentation on two datasets: IFN/ENIT and AHDB Arabic Handwriting Database. Their results showcased accuracies of 97.4% and 98.9%, respectively.

Neche, Chemseddine, et al. (2019) [29] introduced a text-line segmentation approach employing a deep learning architecture. Specifically, they employed an RU-Net enabling pixel-wise classification to differentiate text-line pixels from the background. The experimental assessment was conducted on the KHATT standard Arabic benchmark, and the results obtained validate the successful segmentation process, achieving a rate of 96.7%
.

Gader, Takwa, et al. (2020) [15] developed a system for extracting text lines from images containing unconstrained handwritten Arabic texts sourced from the public Arabic dataset, BADAM. The approach relies on a deep neural network named AR2U-Net, incorporating a Recurrent Residual convolutional neural network in conjunction with the U-Net model and an Attention mechanism. The model demonstrated its performance, achieving a precision rate of 93.2% .

Mechi, Olfa, et al. (2021) [30] introduced a hybrid method that merges a U-Net deep network with traditional document image analysis techniques, including connected component analysis and modified RLSA, to localize text lines in various contemporary datasets of Handwritten Arabic documents, both public and private.

The outcomes demonstrated the efficacy of the proposed approach, achieving a high precision rate of approximately 90% .

Meziani, Fariza, et al. (2021) [31] implemented their proposed technique on document images sourced from the standard KHATT database, which exhibited various inclinations, overlapping, and intersecting lines. Prior to employing the segmentation method, they executed a sequence of preprocessing operations. These steps encompassed the conversion of Gray-scale images into binary ones, employing the Hough transform for skew detection, and rectifying inclinations to ameliorate image quality. To execute the segmentation, they amalgamated three methodologies reliant on horizontal projection profile (HPP), connected components (CC), and skeleton analysis. Their outcomes showcased promise, achieving notable metrics such as an f-measure of 85% .

Gader, Takwa Ben Aïcha, and Afef Kacem Echi (2022) [32] proposed an effective technique for accurately segmenting overlapping and touching handwritten Arabic text lines. Their approach relies on a modified U-Net called AR2U-net, which is a deep learning-based method trained on the LTP (Local Touching Patches) database. This model performs pixel-wise classification to segment touching characters. Additionally, they introduced a post-treatment step to segment consecutive touching text lines, resulting in an impressive accuracy of 94.6% .

Abdo, Hakim A., et al. (2022) [33] their analysis of Arabic text documents introduced a comprehensive four-step methodology: preprocessing, text line segmentation, word segmentation, and character segmentation. The technique leverages horizontal projection methods to detect and extract text lines. In the word segmentation phase, space thresholds are computed to differentiate within-word and between-words spaces, effectively isolating individual words. Following this, a thinning method is employed to identify ligatures and characters. The proposed methodology underwent rigorous testing on a dataset of 115 text images, inclusive of samples from the King Fahd University of Petroleum and Minerals (KFUPM) handwritten Arabic text (KHATT) database, along with additional images generated by the researchers. The experimental outcomes displayed exceptional performance, yielding success rates of 98.6 % for line segmentation, 96% for word segmentation, and 87.1% for character segmentation.

5. Datasets

Most of the experimental studies and research in the realm of automatic segmentation and offline Arabic handwriting recognition rely on diverse Arabic databases. These databases encompass collections of images featuring various content types such as characters, words, numbers, and complete texts. For instance, the AI-ISRA database [14] incorporates Arabic sentences, words, digits, and signatures from 500 individuals. In contrast, the AHDB database [34] encompasses 10,000 words authored by 100 writers, while the IFN/ENIT dataset [18] includes Arabic words and Tunisian town names. The CENPARMI database [35] comprises 3,000 digits (legal and courtesy amounts, and numerals). The IFHCDB database [36] focuses on isolated offline handwritten Farsi/Arabic numbers and characters, showcasing grayscale images of 52,380 characters and 17,740 numerals. AHD-Base [37] comprises 60,000 training digits and 10,000 testing digits scribed by 700 individuals of diverse ages and educational backgrounds. Meanwhile, the AHD/AMSH database [38] features 12,300 Arabic handwritten words produced by 82 writers. Additionally, the Alamri Database [6] is a collection of 46,800 digits, 13,439 numerical strings, 21,426 letters, 11,375 words, and 1,640 special symbols, written by 328 contributors. The APTI (Arabic Printed Text Images) [39] dataset is artificially created using a lexicon consisting of 113,284 words, employing 10 Arabic fonts with various sizes and styles. This database encompasses 45,313,600 individual word images, amounting to over 250 million characters. Conversely, the SUST-ALT database [40] comprises numerals, letters, and Arabic names. The KHATT database [41] comprises 1000 forms and 2000 paragraphs authored by 1000 writers. The HACDB dataset [16] provides a collection of Arabic character images designed to encompass various shapes, including overlapping characters. It encompasses 6600 character shapes created by 50 writers. The AIA9K [42] is a database of the Arabic alphabet, featuring 8737 letters distributed across 28 classes, while the AHCD [8] consists of 16800 isolated Arabic characters. The KU-database [7] is composed of words extracted from renowned Arabic proverbs, encompassing a total of 3024 word images, 14616 PAWs (Part of Arabic Words), and 30744 characters. The recently introduced DBAHD [43] is a proposed database focusing on Arabic handwritten diacritics, covering various forms of diacritical marks. It comprises 500 diacritics distributed across 5 folders, including 100 examples of single-point, double-point, triple-point, Hamza, and madda.

A recent and noteworthy addition is the HAMCDB [44], presented as the inaugural database of handwritten Maghrebi characters, featuring 1560 images. Table 2 provides an overview and summary of these aforementioned datasets.

6. Open Issues, motivation and Future Research Directions

The ancient manuscript heritage represents a big part of the individual and collective memory of the country; it

Table 1
A Summary of the related works.

Author and Year	The segmentation technique	The experiment dataset	Evaluation Metrics
(Boussellaa, Wafa, et al., 2010)[23]	Block covering analysis using unsupervised technique	100 old handwritten document images from the National Library of Tunisia	Accuracy=95%
(Kumar, Jayant, et al., 2010)[24]	A graph-based approach (by building a sparse similarity graph and the use of a shortest path algorithm to compute similarities between non-neighboring components.	125 Arabic document images	Accuracy=96%
(Kumar, Jayant, et al.,2011)[25]	A graph-based technique to detect touching and proximity errors + a refinement operation using Expectation Maximization (EM)	Privet datasets of 125 Arabic document images with 1974 text-lines	F1=98.76%
(Khayyat, Muna, et al., 2012)[26]	A smearing technique based on morphological dilation with a dynamic adaptive mask	CENPARMI Arabic handwritten documents (Section IV-A)	Precision =96.3%
(Al-Dmour, Ayman, and Fares Fraij., 2014)[27]	the horizontal projection profile (HPP)	Benchmarking datasets of the AHDB	extraction rate=84.8%
(Suresha, M., and Amani Ali Ahmed Ali., 2018)[28]	Hough transform approach preceded by a novel method based on skeletonization in the post-processing stage	IFN/ENIT and Arabic Handwriting Database: AHDB	Accuracies=97.4% and 98.9%
(Neche, Chemseddine, et al., 2019)[29]	A deep learning Architecture (RU-net).	KHATT	Accuracy=96.7%
(Gader, Takwa, et al., 2020)[15]	A deep neural network called AR2U-Net based on the U-Net model	BADAM	Precision=93.2%
(Mechi, Olfa, et al., 2021)[30]	Hybrid method (a deep network (U-Net architecture) with classical image analysis techniques).	ANT database	High Precision
(Meziani, Fariza, et al., 2021)[31]	Combination of: Horizontal projection profile (HPP), on connected components (CC) and on skeleton.	100 text images from KHATT	F1=85%
(Gader, Takwa Ben Aïcha, and Afef Kacem Echi, 2022) [32]	Deep learning-based method based on a modified U-Net named AR2U-net (Attention-based Recurrent Residual U-net model).	LTP (Local Touching Patches) database	Accuracy= 94,6%
(Abdo, Hakim A., et al. 2022) [33]	The horizontal projection technique is utilized for detecting text lines, calculating a threshold to determine the spacing between isolated words, and subsequently applying a thinning method to detect characters.	A set of 115 text images from (KFUPM) (KHATT) database + Some images produced by the authors	Success rate= 98.6% (lines segmentation) 96% (words segmentation), and 87.1% for characters segmentation.

has played a main role in the preservation of cultural identity and the construction of the contemporary Arabian states. Despite its importance, it faces difficult situations, some of which result from its transfer from its places of origin, which has caused the destruction and loss of certain rare manuscripts. To safeguard this heritage of ancient texts and manuscripts, custodians and preservationists tasked with its care turn to digital tools and techniques. They organize digitization projects to convert these materials into digital formats, enabling further research through automated procedures. These operations encompass the analysis, access, and comprehension of the manuscript content. Firstly, an artistic card is proposed for each manuscript, containing details ranging from the manuscript's content to its formal aspects. Such information aids in the creation of diverse structured databases across various domains of these manuscripts and facilitates the development of corresponding segmentation and recognition systems.

As a form of motivation and for future research in preserving this significant heritage, we are currently involved in establishing an initial database. This database

Table 2
Summary of well-known Arabic Datasets used in Recognition Systems

The database	Author and year	The data content type
AI-ISRA	(Kharma et al. 1999)[14]	Arabic sentences, words, digits and signatures of 500 writers.
AHDB	(Al-Ma'adeed et al. 2002)[34]	10000 words written by 100 writers.
IFN/ENIT	(Pechwiz et al. 2002)[18]	Arabic words and Tunisian town names.
CENPARMI	(Al-Ohali et al. 2003)[35]	3000 digits.
IFHCDB database	(Mozaffari el al. 2006)[36]	52380 Grayscale images of isolated offline handwritten Farsi/Arabic characters and 17740 numerals.
AHD-Base	(El-Sherif E. A. Abdelazeem S.2007)[37]	60000 digits for training and 10000 digits for testing written by 700 persons
AHD/AMSH	(AL-NASSIRI et al. 2007)[38]	12300 Arabic handwritten words written by 82 writers
Alamri	(Alamri et al. 2008)[6]	46800 digits, 13439 numerical strings, 21426 letters, 11375 words and 1640 special symbols, written by 328 writers
APTI	(Slimane et al. 2009)[39]	45313600 single word images totaling to more than 250 million characters
SUST-ALT	(Musa et al. 2011)[40]	Numerals, letters and Arabic names
KHATT	(Mahmoud et al. 2012)[41]	1000 forms and 2000 paragraphs written by 1000 writers
HACDB	(Lawgali et al. 2013)[16]	6600 shapes of Arabic characters written by 50 writers
AIA9K	(Torki el al. 2014)[42]	8737 Arabic letters with 28 classes.
AHCD	(El-sawy el al. 2017)[8]	16800 isolated Arabic characters
KU-database	(HAFIZ A et al.2016)[7]	3024 Arabic words images, 14616 PAWs, and 30744 characters.
DBAHD	(Lamghari N. and S. Raghay.2021)[43]	500 handwritten Arabic diacritic marks.
HAMCDB	(Soumia D. et al.2022)[44]	1560 images of handwritten Maghrebi Isolated Characters.

will feature scanned and photographed images of ancient Algerian manuscripts collected from diverse centers and regions across the country. It is intended to serve as a foundational resource for various automatic processing experiments, particularly in the domain of text line segmentation.

7. Conclusion

This paper offers an extensive examination of established approaches for offline Arabic text line segmentation and extraction in handwriting. It begins with a concise depiction of the origins and key attributes of ancient handwritten Arabic manuscripts. Subsequently, it explores various techniques employed in segmenting handwritten Arabic documents and delves into the encountered challenges with this script style. Additionally, it provides a comprehensive and comparative analysis of existing works using experimental datasets, serving as a valuable resource for computer vision, machine learning researchers, practitioners, and engineers. Furthermore, it presents an overview of datasets and concludes by addressing unresolved issues, challenges, and future research directions, beneficial for emerging researchers and engineers.

Bibliography

[1] Orsatti P. Le manuscrit islamique: caractéristiques matérielles et typologie. pages 269–331, 1993.

[2] Al-Dmour A Fraij F. Segmenting arabic handwritten documents into text lines and words. In *International journal of Advancements in Computing technology*, page 109, 2014.

[3] Islamic medical manuscripts at the national library of medicine. https://www.nlm.nih.gov/hmd/arabic/arabichome.html. Accessed: 2023-03-10.

[4] Bibliothèque nationale de tunisie. http://www.bibliotheque.nat.tn. Accessed: 2023-03-10.

[5] Lebore T. Segmentation d'image application aux documents anciens. In *Mémoire de Master de recherche, Université de Nante, France.*, 2007.

[6] et al. Huda, Alamri. A novel comprehensive database for arabic off-line handwriting recognition. In *Proceedings of 11th international conference on frontiers in handwriting recognition, ICFHR*, pages 664–669, 2008.

[7] et al Hafiz, A. M. Ku±database of handwritten arabic words. 2016.

[8] El-Sawy A Loey M El-Bakry H. Arabic handwritten characters recognition using convolutional neural network. In *WSEAS Transactions on Computer Research*, pages 11–19, 2017.

[9] A Belaid. Analyse de documents: de l'image à la représentation par les normes de codage. In *Cours de l'INRIA. Document numérique*, pages 21–38, 1997.

[10] Bennasri A Zahour A Taconet B. Extraction des lignes d'un texte manuscrit arabe. In *Vision interface*, pages 42–48, 1999.

[11] Kaileh H. L'accès à distance aux manuscrits arabes numérisés en mode image. In *Doctoral dissertation, Lyon 2*, 2004.

[12] Van den Boogert N. Some notes on maghribi script. 1989.

[13] Wolf C Doermann D. Binarization of low quality text using a markov random field model. In *International Conference on Pattern Recognition, IEEE*, pages 160–163, 2002.

[14] Kharma N Ahmed M Ward R. A new comprehensive database of handwritten arabic words, numbers, and signatures used for ocr testing. In *IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No.99TH8411)*, pages 766–768, 1999.

[15] Gader Takwa B A et Echi Afef K. unconstrained handwritten arabic text-lines segmentation based on ar2u-net. In *17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 349–354, 2020.

[16] Lawgali A Angelova M Bouridane A. Hacdb: Handwritten arabic characters database for automatic character recognition. In *European workshop on visual information processing (EUVIP)*, pages 255–259, 2013.

[17] et al. Papavassiliou, Vassilis. Handwritten document image segmentation into text lines and words. In *Pattern recognition*, pages 369–377, 2010.

[18] Pechwitz M Maddouri S S Märgner V Ellouze N Amiri H. Ifn/enit-database of handwritten arabic words. In *Proc. of CIFED*, pages 127–136, 2002.

[19] et al. Al-Barhamtoshy, Hassanin M. typewritten and handwritten using optical character recognition (ocr) system. In *Arabic calligraphy*.

[20] Ali A Suresha M. Survey on segmentation and recognition of handwritten arabic script. In *SN Computer Science*, pages 1–31, 2020.

[21] Tseng Y H Lee H. Recognition-based handwritten chinese character segmentation using a probabilistic viterbi algorithm. In *Pattern Recognition Letters*, pages 791–806, 1999.

[22] Likforman-Sulem L Zahour A Taconet B. Text line segmentation of historical documents: a survey. In *International Journal of Document Analysis and Recognition (IJDAR)*, pages 123–138, 2007.

[23] Boussellaa W Zahour A Elabed H Benabdelhafid A Alimi A M. Unsupervised block covering analysis for text-line segmentation of arabic ancient handwritten document images. In *20th International Conference on Pattern Recognition*, pages 1929–1932, 2010.

[24] Jayant Kumar, Wael Abd-Almageed, Le Kang, and David Doermann. Handwritten arabic text line segmentation using affinity propagation. In *Proceedings of the 9th IAPR international workshop on document analysis systems*, pages 135–142, 2010.

[25] Kumar J Kang L Doermann D Abd-Almageed W. Segmentation of handwritten textlines in presence of touching components. In *International Conference on Document Analysis and Recognition*, pages 109–113, 2011.

[26] Khayyat M Lam L Suen C Y Yin F Liu C. Arabic handwritten text line extraction by applying an adaptive mask to morphological dilation. In *10th IAPR International Workshop on Document Analysis Systems*, pages 100–104, 2012.

[27] Al-Dmour A. Fraij F. Segmenting arabic handwritten documents into text lines and words. In *International journal of Advancements in Computing technology*, 2015.

[28] Suresha M Ali A. Segmentation of handwritten text lines with touching of line. In *International Journal of Computer Engineering and Applications*, pages 1–12, 2018.

[29] Neche C Belaid A Kacem-Echi A. Arabic handwritten documents segmentation into text-lines and words using deep learning. In *International Conference on Document Analysis and Recognition Workshops (ICDARW)*, pages 19–24, 2019.

[30] Mechi O Mehri M Ingold R Amara N E. Combining deep and ad-hoc solutions to localize text lines in ancient arabic document images. In *25th International Conference on Pattern Recognition (ICPR)*, pages 7759–7766, 2021.

[31] Meziani F Bouchakour L Ghribi K Yahiaoui M Latrache H Abbas M. Arabic handwritten text to line segmentation. In *International Conference on Information Systems and Advanced Technologies (ICISAT)*, pages 1–5, 2021.

[32] Takwa Ben Aïcha Gader and Afef Kacem Echi. Deep learning-based segmentation of connected components in arabic handwritten documents. In *International Conference on Intelligent Systems and Pattern Recognition*, pages 93–106. Springer, 2022.

[33] et al. Abdo, Hakim A. An approach to analysis of arabic text documents into text lines, words, and characters. In *Indones. J. Electr. Eng. Comput*, pages 754–763, 2022.

[34] Al-Ma'adeed S Elliman D Higgins C A. a data

base for arabic handwritten text recognition research. In *Proceedings eighth international workshop on frontiers in handwriting recognition*, pages 485–489, 2002.

[35] Al-Ohali Y Cheriet M Suen C. databases for recognition of handwritten arabic cheques. In *Pattern Recognition*, pages 111–121, 2003.

[36] Mozaffari S Faez K Faradji F Ziaratban M Golzan S. a comprehensive isolated farsi/arabic character database for handwritten ocr research. In *Tenth International Workshop on Frontiers in Handwriting Recognition*, 2006.

[37] El-Sherif E A Abdelazeem S. a two-stage system for arabic handwritten digit recognition tested on a new large database. In *Artificial intelligence and pattern recognition*, pages 237–242, 2007.

[38] Al-Nassiri A ABDULLA S. a new arabic (ahd/amsh) handwritten database. In *ACIT, Lattakia, Syria*, 2007.

[39] Slimane F Ingold R Kanoun S Alimi A M Hennebert J. a new arabic printed text image database and evaluation protocols. In *10th International Conference on Document Analysis and Recognition*, pages 946–950, 2009.

[40] Musa M E. Arabic handwritten datasets for pattern recognition and machine learning. In *5th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–3, 2011.

[41] Mahmoud S Ahmad I Al-Khatib W G Alshayeb M Parvez M T Märgner V et. al. Khatt: An open arabic offline handwritten text database. In *Pattern Recognition*, pages 1096–1112, 2014.

[42] Torki M Hussein M E Elsallamy A Fayyaz M Yaser S. Window-based descriptors for arabic handwritten alphabet recognition: a comparative study on a novel dataset. In *arXiv preprint*, pages 1411–3519, 2014.

[43] Lamghari N Raghay S. Recognition of arabic handwritten diacritics using the new database dbahd. In *Journal of Physics, IOP Publishing.*, page 012023, 2021.

[44] Djaghbellou S. Attia A. Bouziane A. & Akhtar Z. Local features enhancement using deep autoencoder scheme for the recognition of the proposed handwritten arabic-maghrebi characters database. In *Multimedia Tools and Applications*, pages 1–19, 2022.