The Trend of Segmentation for Arabic Handwritten Touching Characters

Ahmed Mansoor Mohsen Algaradi[1], Mohd Sanusi Azmi[2], Intan Ermahani A. Jalil[3]
Abdulwahab Fuad Ayyash Hashim[4], Afrah Abdullah Muhammad Al-Malki[5]
Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka[1, 2, 3]
Network Systems Support Engineering, College of Telecom and Electronics, Jeddah, Kingdom of Saudi Arabia[4]
Computer Science, Umm Al Qura University, Adhm, Kingdom of Saudi Arabia[5]

***Abstract*—The paper is a comprehensive study of existing research trends in the sector of Arabic language, with a focus on state-of-the-art methods to illustrate the existing condition of various theory in that sector, with the goal of facilitating the adaptation and extension of prior ones into new systems and applications. In the Arabic alphabet, there are 28 letters. Depending on its place in the word, every Arabic letter has over one shape; a single character may have from one to four shapes. The Touching between character and the Overlapping occurred in the handwritten. Historical documents contained a massive knowledge and culture. There are many old books that need to be converted into readable format. Which would take a long time if humans converted it. However, the main problem is the lack of research in Arabic Handwritten especially for segmentation of touching characters. Thus, current trends of the segmentation techniques are investigated to identify the current state-of-the art of segmenting touching characters in other domains for constructing enhance techniques for Arabic touching characters. In this paper, it reviewed approaches for the segmentation of the touching characters. This paper presents the trend of approaches for the recognition process and segmentation of Arabic handwritten touching characters. In this paper, it highlighted the strength of each technique, the method used, and the drawback of the techniques. Based on the outcome, this will provide a good foundation for constructing a better technique for segmentation of Arabic touching characters, especially from the degraded documents.**

Keywords—Component; character segmentation; Arabic handwritten; character touching; recognition

I. Introduction

Arabic is now the official language of nearly 26 nations, with a population of 280 million people globally. It is among the six official languages of the United Nations (UN) (Chinese, Arabic, English, French, Russian, and Spanish). Furthermore, several of its vocabulary and forms are used in Persian (Farsi), Jawi, Kurdish, Urdu, and Pashto.

Some individuals here nowadays mostly use pen and paper to write notes (for instance). That strategy has a number of flaws. Handwritten text is difficult to retain and access in an efficient and appropriate manner. Searching through them and sharing them with others is a time-consuming process. A lot of critical knowledge may be lost and not utilized efficiently if that content was not available in electronic form.

The segmentation might confront various difficulties. In addition, character should not be too tinny and neatly segmented to better identify the recognition process [1]. The Arabic word is often a line that draws this intricacy of segmentation [2]. Because it used computers in almost every aspect of life, it also known the modern era as the information technology era. The computer is a necessary component of human life. Although, compared to humans, computers do not have nearly as much intelligence. Humans can recognize any sort of text picture from old and deteriorated texts in libraries, but computers cannot comprehend these text images directly [3] Offline handwritten touching Arabic characters segmentation is a popular topic in study, however it's fraught with difficulties because to differences in writing, overlapping, and touching letters. The segmentation becomes tough when two characters are related to each other [4]. Mostly, all libraries and national archives throughout the world hold large volumes of historical and deteriorating documentation as a book. To convert these important resources to a machine-readable file, special care must be taken [5]. The Arabic language comprises 28 letters, each of which has a distinct form. Because letters in writings are combined to create words, these connections affect the appearance of the letters, thus the shape of an isolated character differs from the shape of a character in the middle and end of the word [6]. Segmentation is closely connected to recognition since it is a highly significant and key phase that splits a picture into sub-units such as lines, words, and letters [7].

OCR (Optical Character Recognition) is a technique that converts scanned or other kinds of pictures into editable format [8]. But even though picture segmentation is not strongly associated with image recognition, the two are inextricably linked. Segmentation process is a critical foundation for image recognition [4]. Picture segmentation, a critical process, splits the picture into tiny pieces.

Even though handwriting is common and varies from person to person, segmentation, which is used to break the text into lines, words, and characters of handwritten text, is still a difficult task. As a result, many observers are going to investigate answers to solve the problem, and some of them have made notable achievements; however, more research is needed to improve the performance of already developed systems. Although it is impossible to explain all the established approaches in this work, the study conducted by addressing the difficulties of touching Arabic handwritten letters [9].

However, this paper aims to show the results and specifications of each segmentation method to assist researchers in determining the best technique for their work.

The rest of the paper is arranged as follows. Section II explains the fundamentals of the Arabic language's characteristics Section III describes the works that are related. Results and discussion details are in Section IV. Section V discusses the conclusion and next work for further study.

II. Related Work

According to a review of the published literature on the segmentation of touching characters, there is a lack of research effort for handwritten and typed Arabic characters when compared to the number of techniques proposed for other languages such as Chinese and English.

In [13], for printed Arabic text, propose a segmentation based on Omni typeface and open-vocabulary OCR. The APTID-MF dataset was chosen as the basis for the suggested approach. This method does not need an explicit font type identification stage. The method used in this work requires cautious management, since picture samples produced by conventional image augmentation algorithms might lose important features and can be linked.

According to [14], to segment Arabic handwritten text, a region-based approach is used to extract the diacritic. After grayscale the picture, they binarize it, then use the region-based method, and finally extract the diacritic from the image. The researcher utilized the Al Quran as a dataset and added 10 handwritten Arabic pictures. This study also addresses diacritics, which are crucial to the syntax and semantics of a word. While it is part of the alphabet, the points and hamza "ء" are considered as diacritics.

Meanwhile [15] the researcher identifies fork points on handwritten Chinese character skeletons. The primary goal of this study is to increase the proportion of segmentation and recognition. The method identifies the feature point in the binary picture, then thins and smooths the character image to identify the fork and endpoints. Following that, they make some changes to eliminate the erroneous branches. They make use of the DHCCCRL database. The rectification of form distortion and the selection of 6,000 handwritten Chinese character pictures are two of the work's highlights.

In addition, [16] method for developing a junction detecting algorithm the researcher omitted a database in this study. This study is just for the Printed Uppercase Alphabet and only between two characters. In this study, just one segmentation instance was investigated. The case presented in this study is neither trustworthy nor practical. In fact, the touching in the writing is more difficult.

Fig. 1 illustrates an instance in which they tested the two characters created by the researcher and placed a straight line between them; normally, it is not touched in this manner during natural handwriting compared to the case shown in the figure below, which is quite significant.

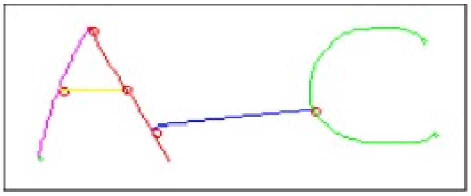

Fig. 1. Example Two Touched Characters [16].

Moreover [17] a technique based on junctions was used to create a handwritten Devanagari character by using a combination of feature to extract the character. Beginning with Handwritten Character Transformation to Bit-mapped Binary Images, the binary image was scaled, and then The Extraction was performed. They get the data from the CVPR Unit, ISI, and Kolkata. One benefit of this research is the collection of 4900 handwritten Devanagari characters. There are five options in this research. On the other side, it might be claimed that there is an advantage to having a lot of options. If, for example, two characters cannot be effectively segmented, the other option can be used.

Furthermore, in [18] Inam Ullah used the junction method while handling Arabic handwritten text. The picture is transformed to binary, and just one point of thickness is kept, making it easier to discern endpoints. However, the intersection set theory is then applied to determine the junction point and broken character. The major goal of this research is to use the algorithm to convert a handwritten, unreadable old Arabic book into a readable one. One advantage of this study is that identifying the endpoints aids in the discovery of the broken character. The researcher chooses the contact point by hand from one of four datasets: IFN/ENIT, CEDAR, and IFN/ENIT, Arabic Dataset, AHDB, and Arabic Handwritten 1.0.

In [19], segmentation of Arabic handwritten text has been performed using contour analysis. In this research, the page is divided into lines initially. Second, the line is divided into sub-words, and last, the sub-words are divided into characters. This method makes use of the database IFN/ENIT. Instead of identifying the baseline or intersecting points, this study replicates the human analogue in Arabic text writing.

Likewise, in Inam Ullah [9], the touching Arabic handwritten characters were segmented using contour tracing. Remove unnecessary noise from a binary picture. Identifying the End, Touching, and Neighboring Points Direction should be written. In the end, they are divided into characters. Many databases were considered, including AHDB, IFN/ENIT, Arabic handwritten 1.0, IBN SINA, IAM, and NIST. Because of proper segmentation, this study could achieve 97.27 percent.

Referring to [20] Corner detection in pictures is a fundamental computer vision problem.

In Lamia Berriche (2020) [24] the technique used is Seam carving-based and Datasets are IESK-ArDB and IFN/ENIT this method leads to Result of 95.67% clear remark for this research is that small characters could be considered secondary components.

Finally, according to [5], the researcher ran one set of 100 words without overlapping and another set of 100 words with overlapping from the benchmark database. And next apply the Method on the handwritten words and report the results for only the second batch. As it stated, it is a simple method that is straightforward to use and quick. Slant correction approaches do not give good results when writing characters with severely slanted and horizontally overlapping characters. Few letters, such as u, v, w, m, and n, are over-segmented or skipped segmented. In Core-zone detection, the researcher advised to count the white pixel until the first major change happens. But how can determine if this one is significant or not? This is a fluid word, and anyone may argue for or against it. Because science only speaks the language of numbers. Their method is straightforward; however, they cannot provide the results of segmentation before and after using the Core-zone detection. As a result, it can be determined if it is essential or not. The researcher simply stated that the first set of words is excellent, with no percentage showing how much is good, so that may compare the overlapping and non-overlapping sets. Also, make it more dependable.

III. Arabic Language Characteristics

A. Location/Direction in Writing

In both handwritten papers and machine printed materials, Arabic text is written from right to left, but numerals are written in the same way as numbers in other languages, for example, from left to right [10], [11].

Fig. 2 shows one example.

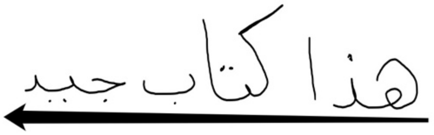

Fig. 2. Direction for Arabic Writing.

Fig. 3 shows an Arabic numeral example.

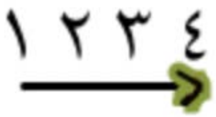

Fig. 3. Arabic Number Direction.

B. Shape of Arabic Characters

Because Arabic writing letters are interconnected with each other, virtually every character in the Arabic language changes its shape in writing in word according to its placement in the word.

Fig. 4 depicts Arabic letters that change shape depending on their location in an Arabic word, as well as instances of how these characters are linked to form words. The picture illustrates four Arabic letters as an example. However, not much different in the rest of the Arabic language letters regarding the shape forms of the letters compared to the selected four alphabets [12].

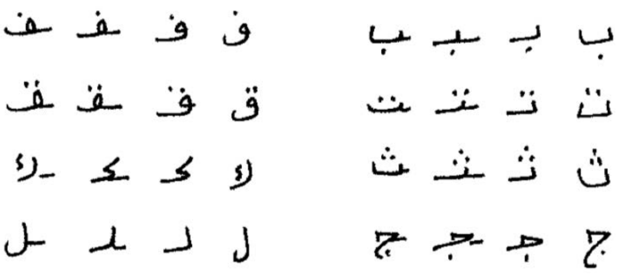

Fig. 4. Example of Shape Alphabet [12].

IV. Challenges and Limitations

There are several obstacles for academics to address in this field, and there is a desire for new ways to develop as computer technology improves and resource constraints diminish [26].

Based on Ouwayed and Belaid's study [23], Kang [22], Aouadi [21], and Saber et al. developed a method for segmenting touching Arabic letters in the same word or other words on the same line or other lines. These existing approaches are template-based segmentation techniques, in which a glossary file is created for all potential touching graphics, that is not only time-consuming due to the variation in Arabic writing and similarities in Arabic characters, but it also fails to address the issue of touching Arabic handwritten characters. Whereas these approaches employed self-defined criteria to govern segmentation accuracy, the segmentation process of touching character pictures suffered as a result.

Over or under segmentation happens because of datasets utilized, languages type (since Arabic has more issues than other languages), type of data (printed or written by hand), and suggested segmentation technique.

By referring to [25] there were some of the challenges such as: Datasets of Arabic handwritten characters, preprocessing noise, Techniques that are cutting-edge, Documents of low resolution and quality, Segmentation, Systems that operate in real time.

The factor that considered as the main factor which is the segmentation. Certain earlier efforts relied on manually dataset segmentation, while others relied on segmented databases. A few of the accessible datasets are not segmented, while others relied on segmented datasets. It's crucial to find a scalable approach to automatically divide documents into lines and subsequently into words (or characters), particularly for big and ancient datasets. Another difficulty in segmentation is dealing with ligatures and the large quantity of Arabic characters.

The multiple sub-words could affect the segmentation process some of the words with single sub-word such as: “محمد” and it could reach to five sub-words for example: “أوروبا” which could increase the difficulties to recognize it as one word during the segmentation process.

Fig. 5 illustrates the challenges of the existing approaches for Arabic Handwritten touched characters segmentation.

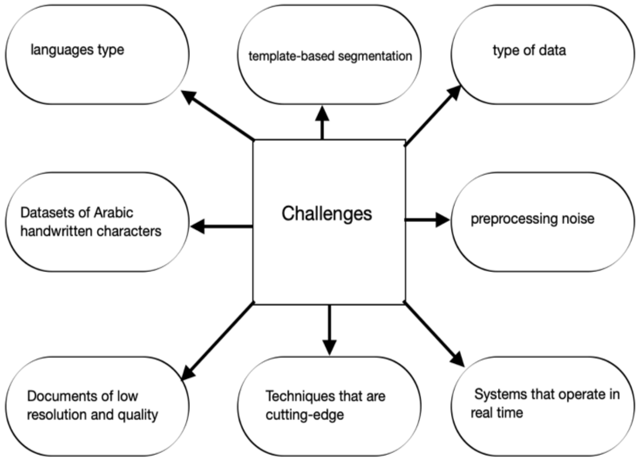
languages type
template-based segmentation
type of data
Datasets of Arabic handwritten characters
Challenges
preprocessing noise
Documents of low resolution and quality
Techniques that are cutting-edge
Systems that operate in real time

Fig. 5. Challenges of the Exisiting Approches.

V. Result and Discussion

After the research method is to find the most successful approach for Arabic handwritten touching character segmentation, but because of the many factors that need to be considered, such as paper quality, number of touching characters that have been tested, database selected, methods used, algorithm applied, and time taken to segment character. Because of all of that, it is difficult to give certain results, especially if some of these factors are not mentioned in the study. However, the author has reviewed ten of the approaches. Table I shows the sample of comparison for each method with its database selected and the result. Author has found a serious need for a specific database which could improve the future research and ease the way for the research to become more reliable, which has a logical result to be compared among the other studies.

TABLE I. Method Comparison

Method	**Methods Comparison**	
	Dataset	***Result***
Seam carving-based	IESK-ArDB and IFN/ENIT	95.66%
segmentation-based on Omni font	APTID-MF	95%
region-based technique	Al Quran	80%
Fork Points on the Skeletons	DHCCCRL	99.41%
junction detection algorithm	-	100%
Junction based approach	CVPR Unit, ISI, and Kolkata	92.8%
junction approach in Arabic	IFN/ENIT, CEDAR, AHDB, Arabic Handwritten 1.0.	93.3%
contour analysis segmentation	IFN/ENIT	89.4%
Contour Tracing	AHDB, IFN/ENIT, Arabic handwritten 1.0, IBN SINA, IAM, NIST	97.27%
Core-Zone	CEDAR	92.6%

The author discovered that the junction algorithm developed by InamUllah yields the highest percentage of segmentation accuracy while being a simple process consisting of three main steps: binary process, thinning process that allows tracing the boundary of the character and if there are more than two binary points, it means there is a junction point to be segmented, and segmentation. However, this study has limitations for future work, such as: during the thinning process, some of the elements may be missing or counted as secondary objects; additionally, the alphabet may be triggered due to its tail. Furthermore, the method could not segment more than one junction point at the same time.

VI. Conclusion

The results of this study revealed current research trends in the field of Arabic. It emphasized the present state of several research elements in that field. This can encourage and make it easier to adapt and extend existing systems to new applications and systems. Arabic has a vast and undiscovered reach; nevertheless, little research has been done in that field previously.

We exhibited some of their prior work that was similar to contemporary state-of-the-art methodologies, with fewer mistakes and a high degree of abstraction. As demonstrated in the difficulties section, this identification is meant to give recommendations for future advancements in the field.

Because of the quality of screening, touching handwritten characters is present in old manuscripts. The Author therefore found that touching characters occurs widely in English, Chinese, Devnagari, Numbers and Arabic handwritten historical materials by exploring the literature for the review. This paper is scanning several approaches to help the researchers in this field to find the advantage and disadvantage of these approaches. For future research, the researcher encourages develop a database for touching characters in Arabic language to give more attention to multiple overlapping.

Acknowledgment

The author would like to thank his brothers Abdulrahman Algaradi and Abdullah Algaradi for sponsoring him during this research.

References

[1] Farulla, G. A., Murru, N., & Rossini, R. (2017). A fuzzy approach to segment touching characters. Expert Systems with Applications, 88, 1-13.

[2] I. Ullah, M. S. Azmi, and M. I. Desa, “Junction point detection and identification of Broken character in touching Arabic Handwritten text using overlapping set theory,” International Journal of Advanced Computer Science and Applications, vol. 10, no. 6, pp. 256–260, 2019, doi: 10.14569/ijacsa.2019.0100636.

[3] S. A. Malik, M. Maqsood, F. Aadil, and M. F. Khan, “An efficient segmentation technique for urdu optical character recognizer (OCR),” in Lecture Notes in Networks and Systems, vol. 70, Springer, 2020, pp. 131–141. doi: 10.1007/978-3-030-12385-7_11.

[4] Farulla, G. A., Murru, N., & Rossini, R. (2017). A fuzzy approach to segment touching characters. Expert Systems with Applications, 88, 1-13.

[5] Saba, T., Rehman, A., & Zahrani, S. A. (2014). Character segmentation in overlapped script using benchmark database. Computers, automatic control, signal processing and systems science, 140-143.

[6] I. Kacem, P. Laroche, Z. Róka, Institute of Electrical and Electronics Engineers. French Section, O. et M. des S. Université de Lorraine. Laboratoire de Conception, and Institute of Electrical and Electronics Engineers, 2014 International Conference on Control, Decision and Information Technologies (CoDIT) : proceedings : Université de Lorraine, France, LCOMS, Metz, November 3-5, 2014.

[7] A. Lawgali, "A Survey on Arabic Character Recognition," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 8, no. 2, pp. 401–426, Feb. 2015, doi: 10.14257/ijsip.2015.8.2.37.

[8] N. Vincent and J. M. Ogier, "Shall deep learning be the mandatory future of document analysis problems?," Pattern Recognition, vol. 86, pp. 281–289, Feb. 2019, doi: 10.1016/j.patcog.2018.09.010.

[9] I. Ullah, M. S. Azmi, M. I. Desa, and Y. M. Alomari, "Segmentation of touching Arabic characters in Handwritten documents by overlapping set theory and contour tracing," International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 155–160, 2019, doi: 10.14569/ijacsa.2019.0100519.

[10] Khreisat, L. (2006). Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study. DMIN, 2006, 78-82.

[11] A. Amin, "OFF-LINE ARABIC CHARACTER RECOGNITION: THE STATE OF THE ART Arabic characters Off-line recognition Handwriting recognition Segmentation Feature extraction Neural Network classifiers Hidden Markov Models Optical character recognition," 1998.

[12] Y. Boulid, A. Souhar, and M. Y. Elkettani, "Handwritten Character Recognition Based on the Specificity and the Singularity of the Arabic Language," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 4, no. 4, p. 45, 2017, doi: 10.9781/ijimai.2017.446.

[13] A. Qaroush, A. Awad, M. Modallal, and M. Ziq, "Segmentation-based, omnifont printed Arabic character recognition without font identification," Journal of King Saud University - Computer and Information Sciences, 2020, doi: 10.1016/j.jksuci.2020.10.001.

[14] A. A. Sheikh, M. S. Azmi, M. A. Aziz, M. N. Al-Mhiqani, and S. S. Bafjaish, "Diacritic segmentation technique for Arabic handwritten using region-based," Indonesian Journal of Electrical Engineering and Computer Science, vol. 18, no. 1, pp. 778–784, Jan. 2020, doi: 10.11591/ijeecs.v18.i1.pp478-484.

[15] Liu, K., Huang, Y. S., & Suen, C. Y. (1999). Identification of fork points on the skeletons of handwritten Chinese characters. IEEE transactions on pattern analysis and machine intelligence, 21(10), 1095-1100.

[16] U. K. S. Jayarathna and G. E. M. D. C. Bandara, "A Junction Based Segmentation Algorithm for Offline Handwritten Connected Character Segmentation," 2006.

[17] IEEE Region 10. Colloquium (3rd : 2008 : Indian Institute of Technology Kharagpur), Institute of Electrical and Electronics Engineers. Kharagpur Section., IEEE Sri Lanka Section., and Damodar Valley Corporation., IEEE Region 10 Colloquium and Third International Conference on Industrial and Information Systems : ICIIS-2008, December 8-10, 2008 : theme: "Real-time communicative intelligence for tomorrow's industry" : e-proceedings. IEEE, 2008.

[18] I. Ullah, M. S. Azmi, and M. I. Desa, "Junction point detection and identification of Broken character in touching Arabic Handwritten text using overlapping set theory," International Journal of Advanced Computer Science and Applications, vol. 10, no. 6, pp. 256–260, 2019, doi: 10.14569/ijacsa.2019.0100636.

[19] Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on. [publisher not identified], 2013.

[20] J. Zhang, T. Luo, G. Gao, and L. Lian, "Junction point detection algorithm for SAR image," International Journal of Antennas and Propagation, vol. 2013, 2013, doi: 10.1155/2013/357379.

[21] Aouadi, N., Kacem, A., and Belaıad, A., 2014. Segmentation of touching component in Arabic manuscripts. In Proceedings of the ICFHR, 1(4), pp. 452–457.

[22] Kang, L., Doermann, D. S., Cao, H., Prasad, R., and Natarajan, P., 2012. Local segmentation of touching characters using contour based shape decomposition. In 2012 10th IAPR International Workshop on Document Analysis Systems, pp. 460–464.

[23] Ouwayed, N., and Belaïd, A., 2009. Separation of overlapping and touching lines within handwritten arabic documents. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5702 LNCS, pp. 237–244.

[24] Berriche, L., & Al-Mutairy, A. (2020). Seam carving-based Arabic handwritten sub-word segmentation. Cogent Engineering, 7(1), 1769315.

[25] Balaha, H. M., Ali, H. A., & Badawy, M. (2021). Automatic recognition of handwritten Arabic characters: a comprehensive review. Neural Computing and Applications, 33(7), 3011-3034.

[26] Eikvil, L. (1993). OCR-optical character recognition.