Characters Segmentation from Arabic Handwritten Document Images: Hybrid Approach

Omar Ali Boraik[1], M. Ravikumar[2]
Department of Computer Science
Kuvempu University
Shankaraghatta, Shimoga 577451
Karnataka, INDIA

Mufeed Ahmed Naji Saif[3]
Department of Computer Applications
Sri Jayachamarajendra College of Engineering
Visvesvaraya Technological University (VTU)
Mysore, Karnataka, INDIA

***Abstract*—Character segmentation in Unconstrained Arabic handwriting is a complex and challenging task due to the overlapping and touching of words or letters. Such issues have not been widely investigated in the literature. Addressing these issues in the segmentation stage reduces errors in the segmentation process, which plays a significant role in enhancing the accuracy of the Arabic optical character recognition. Therefore, this paper proposes a hybrid approach to improve the accuracy for interconnection, overlapping or touching character segmentation. The proposed method includes several stages: removing extra shapes such as signatures from the document. Using morphological operations, connected components and bounding box detection, detect and extract individual words directly from the document. Finally, the touching characters segmentation is achieved based on background thinning and computational analysis of the word's region. The proposed method has been tested on KHATT, IFN/ENIT database and our own collected dataset. The experimental results showed that the proposed method obtained high performance and improved the accuracy compared to other methods.**

Keywords*—Arabic handwritten character recognition; connected components; word segmentation; character segmentation; morphological operators; overlapping and touching characters

I. Introduction

Recently, electronic devices and modern technology have become an important and essential in human daily life. A lot of efforts and time were spent to protect and maintain valuable historical documents, letters, and books into digital images for scientific, service, and future uses. Optical recognition systems appeared as significant tools to avoid the loss of such valuable documents which convert text images into editable digital texts. There are various uses of machine learning techniques in Optical Character Recognition (OCR) systems; the gap is still largely due to unlimited obstacles in Arabic handwriting. Comparing OCR to Latin and other languages with the recognition of Arabic, the Arabic recognition system is still incomplete and unsatisfactory. In today's world, interested parties in the field of documenting are required to save digital images possible to modify, i.e., repairing deterioration in historical books, correcting errors, using a text part of it in other applications.

Converting text images into editable digital forms is called Optical Character Segmentation (OCR) [1]. The text images are either printed or handwritten. The deficiencies of Arabic handwriting OCR systems are more complex than printed and incomplete. Because handwriting does not abide by the font's criteria, specific size, different font and size style writing for a word repeated several times in the same document itself, other problems of interconnection, overlapping and touching, and difference gaps among word/sub-words increase the complexity of Arabic handwriting. There are common factors that make Arabic handwritten as well as printed text complex such as Arabic nature cursive, writing from right to left, connecting the letters, and so on.

Arabic handwritten character recognition has the same situation as other languages. In some cases, it seems to be more complex depending on the language, the challenges it faces, and difficulties for line, words [2]-[4] and the challenges in character segmentation of the input document images. These challenges, such as interconnection letters word, cursive overlapping, touching existence of ligatures, diacritical marks and the position and number of dots above or under some letters. These challenges may lead to misclassification and unsatisfactory results. However, some of these challenges were recognized by many researchers in the OCR field using machine learning techniques [2].

The deficiencies are based on the failure of the Arabic character segmentation stage. Cursive, overlapping, and unrestricted writing challenges are the most long-term barriers to correct segmentation. A study [1] presented a projection profile technique for Arabic characters segmentation; which was tested and evaluated successfully in Arabic database with various sizes, styles, and font types. But it is limited to the Arabic printed document. The proposed system fails to deal with handwriting documents.

Another study [4] presented a solution for overlapping and touching Arabic characters segmentation by overlapping set theory and contour tracing. It is low accurate when segmenting the multiple touching letters. While [6] suggested a hybrid method focusing on the middle point of the word area. This study was focused on handwriting documents in the Hindi language. This method succeeded in fragmenting the multi-touching letters, and to apply it to the Arabic language; it needs to be developed because the Arabic calligraphy starts from right to left.

The segmentation solutions for single/ multiple touching, overlapping challenges, and interconnected characters problems still need to be expanded to include more Arabic handwritten documents. Achieving progress in respect to Arabic text recognition is hindered by such obstacles and challenges. Moreover, the complexity of finding a solution to the segmentation of overlapping and touching Arabic handwritten words or characters made few researchers interested in addressing these two problems and developing techniques to address these complex problems. These two challenges created gaps in attempts to process them, i.e., there is a gap of at least two years between every two research works (Ouwayed et al., 2009, Belaïd & Ouwayed, 2011, Aouadi, N. et al., 2013, Aouadi & Kacem, 2017, Ullah et al., 2019, AbdAllah, N., & Viriri, S. 2021, Ahmed et al., 2022 as a review survey). These are some of the motivations that encouraged us to propose a hybrid approach to reduce these complexities and improve segmentation techniques.

This article proposed hybrid approach to segment the Arabic handwritten document into direct words with an objective to enhance and increase the accuracy ratio of Arabic handwritten character segmentation while dealing with overlapping, interconnected, and touching Arabic handwritten documents. The work was divided into eliminating non-textual appendages in documents, segmentation, and extracting words from the image using connected components and thinning techniques. A hybrid method is used with computation analysis of the word's region to segment a word into characters. The middle point is detected to extract the structure features for dealing with the input, which contains the touching overlapping character and distinguishes from isolated character based on calculating the vacant space index value. The hybrid method has been proven to be a flexible and efficient method to deal with various renewed database. The contributions of this study are listed below:

1) Proposing a hybrid approach for character segmentation in Arabic handwriting that addresses the challenges of segmentation of touching characters eliminates the signatures from input document images and segments words directly from the input images, including wavy lines.

2) Creating a new database for Arabic handwriting to evaluate the proposed hybrid approach.

This article is organized as follows: Section 2 reviews the existing studies related to Arabic characters, Section 3 describes the methodology for character segmentation, Section 4 discusses the experimental results, and Section 5 concludes the study.

II. Related Work

The challenges of touching and overlapping lead the researchers to focus and work on line/ word/ character segmentation. The OCR systems depend on hybrid techniques, which are considered better than other systems in accuracy, speed, and flexibility in dealing with renewable databases. It is appropriate for dealing with handwritten documents as indicated in [7]. The researchers suggested a method in which they used a technique based on connected components. Then, it selects the estimating and the alignment transformation of these connected components, stored in segmented models by templates for two most similar Connected Components (CCs) of the touching words. Finally, it separated the CCs into two regions using the Centre Points of these regions.

The authors [4] introduced a solution for touching Arabic characters by using a hybrid approach for character segmentation. The proposed method to split through letters is by selecting the touching point by overlap set of theory and endpoints of Arabic word and then tracing the boundaries of the touching letters to segment them individually. The segmentation is good, just dealing with single touching characters. Due to the differences in positions of words in different documents, the hybrid methods need to be improved and developed to give better results in the future. The authors [8] suggested a method for touching Arabic characters segmentation based on template segmentation. This method creates a dictionary file. This file contains a template for each touchpoint with its necessary details. Then it is compared with the input images to identify a sample from the dictionary file. As mentioned earlier, the template method is difficult to apply in handwritten character segmentation. It saps computer hardware to save each touchpoint in a template for detecting similar touching points. The runtime is very long and tedious. Fatal flaws in the handwritten character segmentation.

In the study [9], the authors used a segmentation method that extracts topographic features. These features identify potential segmentation points of the characters block connected. The segment of the possible points is based on the average width of the character. The study achieved approximately 70% of character segmentation. Moreover, many errors in segmenting, handwritten Arabic characters, mainly the letters which connect from two sides. While in [10] suggested a method to segment touching handwritten Arabic characters. It first detects the intersection points and the beginning ligature's pixel near the upper line from the baseline. The process starts from this point (ligature's pixel) to pursue the descending character to the intersection point and respect a different angular corresponding to the descending character curvature. The proposed method was tested on 100 Arabic document images containing 256 touching lines. However, the success ratio reached 94% of the segmentation. But it is inappropriate for the segmentation of multiple touching components. Three methods from the literature were compiled and already developed. So, it is a mixture method. The hybrid method is suggested by [11] to segment the touched printed and handwritten Latin characters (obviously, the success of segmentation for handwritten characters. will be the highest efficiency and accuracy in printed characters). The disadvantage of this method is that it was tested on a few images of documents. Therefore, the success rate is insufficient and unsatisfactory.

The problems of overlapping, interconnection, and touching in handwritten Arabic characters make the segmentation process complicated task. From the aforesaid literature review, it was noted that these issues had been extremely slightly discussed in previous studies. In addition, period intervals between those studies were absolutely longer. However, there are shortcomings in solving these problems, which involves limitation in the studies related to Arabic

databases and lack in performance of segmentation of the multiple-touching Arabic characters. The proposed methods are suitable for the segmentation process but are inflexible in dealing with morphological differences. Therefore, this study aims at developing a hybrid method to address the above-mentioned problems.

III. Proposed Method

This section describes the methodology of the proposed work as shown in Fig. 1. Initially, the pre-processing is performed for the input image to improve low quality and prepare for the next stage. Then, it segments the whole document into individual words, especially if the input image contains wavy lines. If the input image includes shapes or signatures, the approximate polygon methods remove these shapes. In contrast, the signatures are extracted and removed from the document based on the connected component analysis. Finally, divide the words into isolated letters.

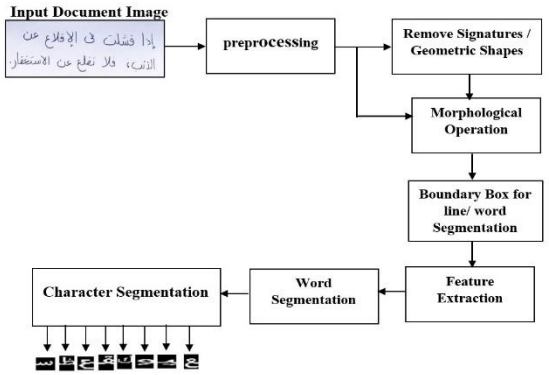


Fig. 1. Block Diagram of the Proposed Method.

A. Pre-processing

In this work, a new database is created which contains old and modern handwritten documents. The quality of the collected images is good, medium, and many of them lack lighting. The problem of poor and unstable lighting in the whole document is big obstacle to make the pre-processing stage complicated as shown in Fig. 2, where the different spots of lighting in same (one) document, which lead to a lack in performing the segmentation process. Using some filters such as Gaussian Filter, Median Filter, Low pass Filter, Custom Filter, etc. which give enhancement in these images, but the results are poor because the most challenge was the unstable lighting in the image documents in which one image has different contrast lighting. The results, on the other hand, were good when the input images with stable lighting were applied which is similar to applying IFN/ENIT or KHTT database, because these databases have Binarization images.

One of the challenges faced when colleting the images, some of the images have poor quality and resolution as shown in Fig. 2. The image (a) is darker than others. Its histogram in (b) placed at 0 - 200 to represent (the) pixel value in x-axis takes the image (a) pixel value and the value placed at 0 to 200 that indicated the values grouped at black values, this represents the image (a) consisting of more black pixels compared to other gray level value. The image (c) is a low contrast image, its histogram (d) is placed at almost the center towards white. The image (e) is a bright image where its histogram (f) values are placed at 170 towards white 255, which indicated the image (e) has more white pixels. Although the preprocessing approaches were applied to these images the results were not good.

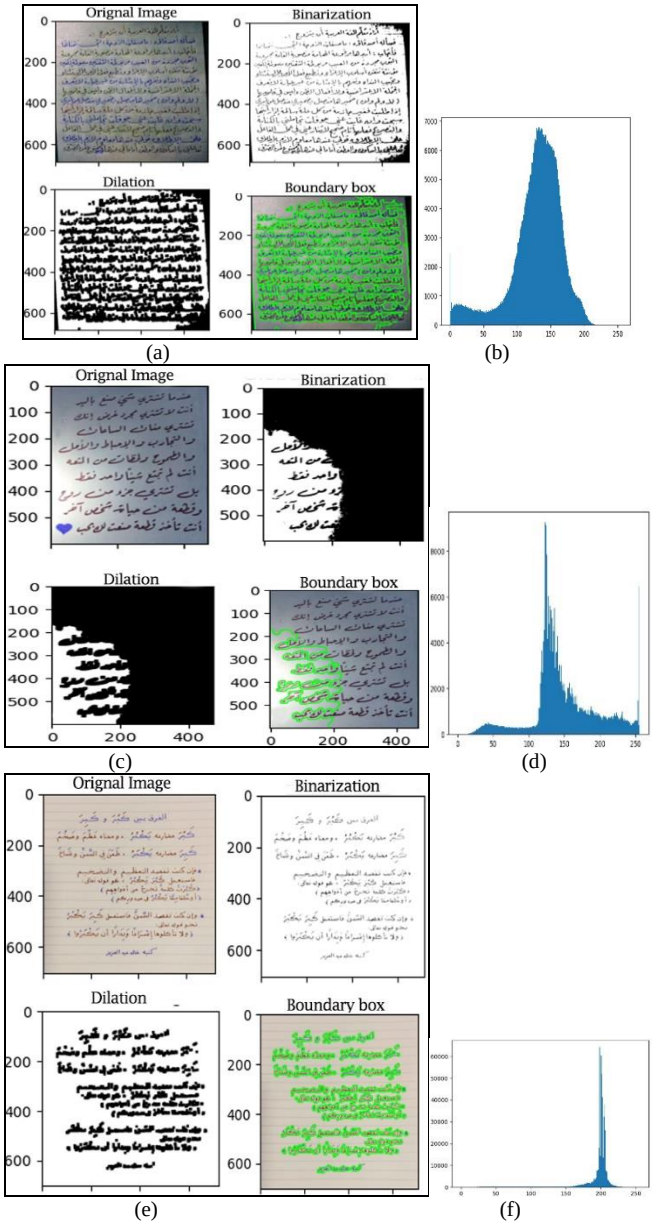


Fig. 2. Different Lighting and Histogram of Intensity Distribution. (a), (c) and (e) are the Input Document Images, While (b), (d) and (f) are the Original Images' Intensity Distribution Histogram for showing the Intuition about the Contrast Brightness Intensity Distribution.

B. Removing the Signatures

Many images of documents contain signatures, so before shifting to the next stages of Arabic OCR stages, it should eliminate these signatures to help to word and character segmentation. The method is used to remove these signatures depending on Connected Components and find out the thresholding average as shown in Fig. 2. By using image processing, the regions of connected pixels were recognized by this algorithm. It commonly gives the same result. In other words, the given input image is scanned by these connected components along with this attached signature. The next step is gathering the pixels into new components connectivity i.e. the elements of the image connected to the same intensity values of each pixel and showing the link with other values. Therefore, the components will be recognized, and every pixel will be highlighted with a specific colour (color labelling). Each pixel may highlight with a grayscale according to the located component.

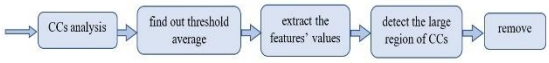


Fig. 3. Block Diagram for removing the Signatures.

Now-a-days, classification of each connected component along with the assorted dissociates are essential to many analysis applications of image's machine-driven. In this process, the whole image, from left to right and top to bottom, is scanned to recognize the region of each pixel which is connected to the image's component. In other words, it can be said the adjacent pixels of each component share constant value V. it can be applied to binary or grayscale images, and it measures connectivity differently. Before applying the mask technique move over, the input image should be a binary, 8-connectivity where a mask created and each pixel and its surroundings is are checked using this 8-connectivity. The operator moves over the image to scan rows individually until it arrives at 'p' point, It examines the remaining eight neighbours of the labelled pixel (at any stage, 'p' is the labelled pixel) for which V = [1]. Then it examines the four neighboured labelled pixels from right to left and from the upper diagonal direction, which were already encountered in the scan. According to previous details of the scan, the term 'p' is classified when the process finds an adjacent value equal to '1'. Then the label is assigned to 'p'. At the same time, if other neighbours have the same value, all of the labels will be assigned. The equivalencies will be noted; if all pixels are 0, a new label will be given to 'p'. This process is followed by the initial scan of the label pair area units and sorted into equivalent categories (classes and distinct labels). The next step is the second scan of the image in which every label is replaced by the equivalent category, even though the labels may not be identical.

A Scikit image library provides an exciting feature to identify and label the connected components. This library is used to check the scanned input image documents and find these connected components with their labels in addition to grey and color labels. It turns out that the largest connected component is the signature compared with word components. Therefore, if it is possible to extract the largest component of the whole document, the signature can be recognized. However, using large connected components can extract unwanted words, lines, or other shapes. Therefore, a threshold value is used to solve this problem, it is used to detect outliers, i.e., any lines, structures, and texts that do not belong to signatures are calculated after a series of experiments which have been performed. In terms of a mathematical formula (1), which is obtained based on experimental results. It gives quite effective recognition of signature's regions in dealing with most A4 size scanned documents. Table I shows the obtained values representing the signature region's characteristics. The place of the signature is determined based on the biggest of the connected component's value and then compared with the average of this region to extract the signature to delete. Fig. 4 shows the identification of the signature's region from a sample set of images.

$$\text{const\_A4} = [100 + (250 * (\text{average}/84))] \quad (1)$$

TABLE I. FEATURES OF SIGNATURE' REGIONS

Sample Images	Biggest Components	Average	Small Components	Big Components
A	924	162.508772	583.657059	10505.82707
B	7682	793.734940	2462.30634	44321.51463
C	1158	128.257353	481.718312	8670.929621
D	1029	38.2051282	213.705739	3846.703298
E	1156	77.5617021	330.838392	5955.091185
F	5888	123.899471	468.748425	8437.471655
G	566	85.7945946	355.341055	6396.138996

Fig. 4. Extraction of Signatures from Sample of Images.

C. Line Segmentation

Separating the input images into isolated lines is successful with high accuracy in documents which may contain printed, historical or handwritten texts, these documents' lines have gaps between them that are orderly spaced [12], and the handwriting is on lined sheets. These gaps are ordered and

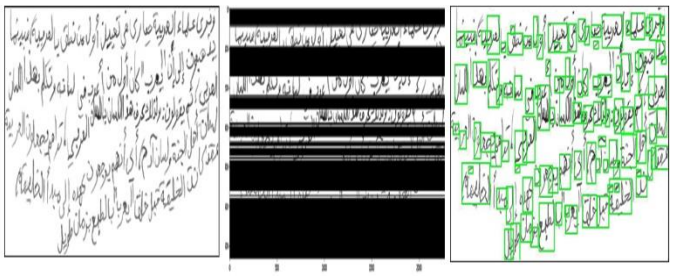
these gaps may be almost empty of noises or unimportant objects. Because the feature of the gaps between the lines is the most important for the success of lines segmentation and extraction individually for using another application, in skew correction with multiple slanting or copy particular lines to other programs. Many handwritten documents lack clear orderly gaps between the lines, which are close, touching, and sometimes wavy. These gaps may be spaced at one side and touching or overlapped at the other. These challenges make the line segmentation process difficult and complex. Many studies have proposed techniques to solve these challenges and achieve satisfying success despite the convergence, contact and overlap between the lines. A study [13] proposed an A* path planning algorithm to line segmentation directly. The localization of each line is detected by two steps: Binarization method and projection profile analysis. The author's experiments tested handwritten historical documents from the MONK and the Saint Gall dataset. The study [14] presented a method of line segmentation based oriented anisotropic Gaussian Kernel, in which the authors divided the input text image into connected components achieved by Boundary Boxes. The method was tested on a sample of English Handwritten documents. In [15], the authors used midpoint detection technique for every two lines or words. The technique was tested on Gurmukhi printed and handwritten (Handwritten scripts are orderly distances among the lines) scripts (Punjabi). Their results achieved 95% of success. Study [5] also used a technique and a database containing English historical documents. The line regions were detected using rough detection throughout black/white transition map which is used to extract the lines through corresponding lines axes and skeletonization. PROHIST database was used in this study and achieved 82.18% of accuracy.

After it converted the input image into binarization in preprocessing phase, the input image locates the text (object) areas of the white background where the text is black. The process Uses Dilation operations to make the pixels' value of each line. This pixels' value has a single value so the line is considered one-component [16]. The starting and ending four points of each component determine the bounding box location among these points around the line component [17]. This boarded rectangle is segmented for each line, returned the segmented line to its original size using the erosion operation and the thinning method to give a better distance between words in preparation for their segmentation. Then the matrix of the rectangle border is saved into an image.

The overlapping and touching challenges between lines, make the dilation operation more extended, two or more lines are considered one component [18]. These challenges are solved by tracing the contour points of the touching area horizontally and overlapping vertically, using the method (direction contour tracing) in [18]. It traces the overlap path with calculation operations. This method is not good in separating lines in Arabic Handwritten documents because of the stretching of strokes and some letters are written vertically and extendly, such as these letters ‘أ’,‘ك’,‘ظ’,‘ط’,‘م’. Also, most line segmentation techniques fail to segment wavy lines.

D. Word Segmentation

As it was mentioned earlier, many of the input images contain lines skewness, touching, overlapping, wavy and more closely spaced lines, where the line extraction fails, so the lines are separated into two or more lines as shown in Fig. 4. In the study [13], Hidden Markov models were used for word segmentation from the entire document directly. Current study, the words or sub-words are segmented by applying the Connected Components method for overcoming and solving many challenges such as the distance between words is less than that between sub-words and overlapping in one word. The Boundary Box function is used to extract them. These two methods achieved a higher success rate and are better than dividing the lines before the words, as shown in Table III.



(a) Original Image. (b) Bad line segmentation. (c) Words segmentation.

Fig. 5. Segment the Document Image into Words.

Fig. 5(a) shows a sample of a document image which contains wavy and touching lines. Many methods were applied to segment the lines in such image. The line segmentation methods such as projection profile method, a method based on tracing, another method based on contours, third method based on baseline, forth method based on morphological operation or other methods. These methods did not succeed in extracting the lines as shown in Fig. 4(b), where the first line only was successfully extracted. The rest of the lines were considered one component, or extracted a part of a particular line with previous or next line in the same process. Fig. 4(c) shows the words were extracted directly from the document with better accuracy and success.

E. Character Segmentation

This step is crucial to the segmentation stage. After overcoming the previously mentioned, challenges as much as possible, such as different light spots in the input image and existing shapes or signatures. After the success of words/sub-words segmentation, the character segmentation follows.

First, detecting the touching of characters: The Connected Components method is used to solve the overlapping problem between characters. It is also used to measure the weight of the character. A threshold value has been fixed to evaluate the weight. If the obtained value is less than the threshold, it can be considered a single character and split automatically. If the value is greater, it can be regarded as a touching character using the variable:

Tc (2) is the aspect ratio of the touching character is greater than the character is automatically split. In order to determine the touching characters using the variable (Tc), this aspect needs to be improved due to the similarity between interconnected and touching of Arabic characters. The touching characters are defined by.

$$T_c = \frac{e^g}{1+e^g} \quad (2)$$

Where $g = \frac{w}{h}$, w is width of the character, and h is height. After identifying the touching character, it is classified as either horizontal, vertical or multiple touching.

By comparing the two values of height and width, the type of touching is determined by $g = h1 > w1$ the touching is vertical. $g = w1 > h1$ the touching is horizontal. While the multiple touching is defined as $g = w2h2 > w1h1$.

Second, using hybrid approach by following these steps:

1) Find the area of a word/ sub-word throughout detect the start point and endpoint of this area, allocate the area on width (w) and height (h). (w, h = contour area(word)).

2) Using thinning, closing and opening operation

3) Using midpoint steps to separate word/ sub-word isolated.

- The vacant space index value is calculated on the word/ sub-word's constrained such as width and height.
- Check the previous pixel and the next one (i+1, i+2, i-1, i-2) to save the column values and check if broken character appears.

4) Saved the vacant space between the characters in array_index.

- Detect the centre value of every vacant space between the characters until the end of the word.

M_point = (start index + last index)/2.

- This mid-point as a centre value is considered as the detection points to split the isolated character.

5) To determine the existence of joining (touching) characters, the total number of characters in a word is calculated.

6) Total value of characters is detected by the ratio of width and height.

Total = $\frac{width}{height}$

7) Total value of characters is compared with the segmentation point.

Value of M_point = total + 1.

8) If joining character is exist then number of segmentation points does not exceed the total no of characters in a single word. Otherwise go to f.

9) f- Calculate the distance sequentially between the middle values, if the distance above 110% of height, there must be a joining character present, which could be single or multiple joining characters.

*10)*Using clustering method to find the cluster in identified area of importance of the character in the middle part.

*11)*Discover the region of importance cluster between M_point1 + 10-(M_point2 +10) to obtain the heap of pixel.

- Scan every column to determine the cluster, if pixel calculate is found to be 10 then it is considered as joining point of the character.
- By leaving three columns in a row, you can segment the joining character.
- The new segmentation points should be extracted.
- Split the word from all the segmentation points.
- Show the results as in Fig. 6.

Fig. 6. Sample Images Show the Character Segmentation Isolated, Touching and Joined Characters.

IV. Evaluation Metrics

Evaluate the performance of the proposed hybrid approach for character segmentation. The performance evaluation metrics used by [19] are followed; the same evaluation strategy was also applied in this work, which uses five factors for evaluation: successful segmentation rate (SR), precision (P), recall (Re), correct segmentation rate (CS), and F-measure (Fm). The authors are illustrated in Equations 3-7. These factors are figured out by counting the number of matches between the resultant segmented words and then characters by the algorithm and ground truth characters in text word segments.

$$SR = \frac{NCc}{NCr} * 100 \quad (3)$$

$$P = \frac{(NCc+NCo)}{NCr} \quad (4)$$

$$Re = \frac{(NCc+NCo)}{NCg} \quad (5)$$

$$CS = \frac{NCr-(NCi+NCo)}{NCg} \quad (6)$$

$$F_m = \frac{(Re*P)}{(Re+P)} * 2 \quad (7)$$

Where NCg: Number of ground truth words, characters respectively.

NCc: Number of segment correct words, character isolated.

NCr: Number of segmentation results (words, characters).

NCi: Number of incorrect segmentations for words, characters.

NCo: Number of over segmentation of touching in words, characters.

V. Experimental Result and Discussion

This section describes the implementation and evaluation of the proposed method. The proposed method was implemented using Python 3.8.8, Open CV environment (Spyder4[MSC V.1916 64 bits]), Win 11 pro 64_bit OS, with Intel(R) Core (TM) i5-9300HF CPU 2.40 GHz, RAM 8 GB.

The proposed method is tested on three Arabic databases. The images of the first database that have been collected were 2,300 handwritten text images, 20156 lines, 302,348 words; the ground-truth value of lines and words were calculated manually and estimated. The collected images are obtained from a scanner and also through social media. These images contain shapes, graphics and signatures. Some of these images are poor in quality and lighting. In such cases, the pre-processing methods such as filters, binarization morphological operation, normalization, skew correction were applied for image purification. The shapes were removed based on the size of the shape region while the signatures were removed from the document images. 140 images contain signatures as shown in Fig. 3. Removing the signatures process was successful in 122 images, the process failed in 7 images, over-segmentation of signatures was noticed in 11 images. However, features of any word's region are similar to features of signature's region results in being removed as shown in earlier Fig. 3.

The proposed method failed in 170 images with a failure rate of 7% due to the poor quality of the images sent through social media. 15,721 lines were segmented correctly. Table IV shows the results of line, word, sub-words or characters segmentation. Because lines are touched, closed, wavy, or slant, the words, and sub-words were directly extracted from the input image. Arabic word's letters may be interconnected overlapped, so the segmentation process of a word depends on the gaps between every two words or sub-word. Table II and Fig. 6 show the experiment results.

KHATT database contains 1000 images of Arabic handwritten forms written by 1000 writers from different countries, 9327 lines, 165890 words [20]. And IFN/ENIT database involves 26459 words which are names of Tunisian cities [21]. Testing the proposed model on three databases was in stages: First, constant documents were segmented into lines, while inconstant documents contain overlap, wavy, or touching problems which are mostly segmented into direct words. The hybrid approach achieved high success rates, as shown in Table II which figures out the distribution of success rates over the three databases. The proposed method achieved a lower accuracy rate in our database for the following reasons:

1) Poor lighting and quality problems.
2) Degradations in the document images.
3) Colored images, while the images in the other two databases are binary.
4) Our database involves more documents than the other two databases.

Table II and Fig. 7 show the words and character segmentation for three databases. In addition, 1450 Arabic handwritten document images were taken from the three mentioned databases. This study focused on images containing overlap and touch problems of words or characters and wavy lines. Table III illustrates the results of the proposed method.

TABLE II. Segmentation Accuracy of Words and Characters

Databases	Number of:		Correct Segmentation		Incorrect Segmentation		Accuracy Rate	
	Words	Characters	Words	Characters	Words	Characters	Words	Characters
KHATT	165890	589924	161411	548629	4479	41294	97.3%	93%
IFN/ENIT	26459	212211	24342	188867	2910	12732	92%	89%
Our DB	302348	1257191	275136	1111356	27211	150863	91%	88%

TABLE III. Segmentation Accuracy of Touching Words and Characters

Inputs	NCg	NCr	NCo	NCi	NCc	SR	P	Re	Fm	CR
words	26468	25409	985	833	25357	99	1.00	0.99	1	89.130
characters	79404	74639	2896	2328	74071	99.23	1.00	0.96	0.99	87.4

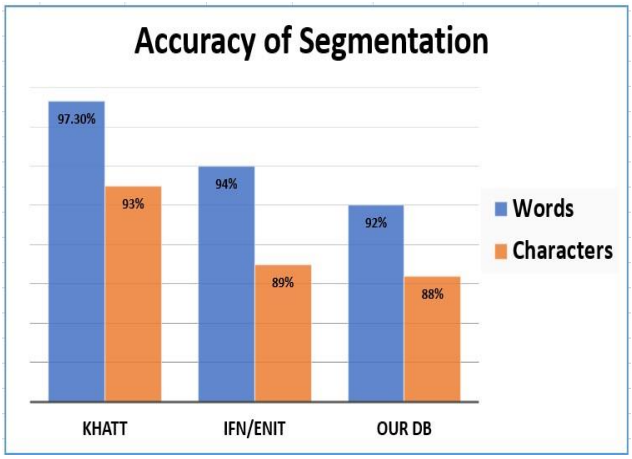


Fig. 7. Segmentation Accuracy of Words and Characters for each Database.

Second: 26468 Arabic handwritten words were taken from the three databases, which contain complex touching and overlapping problems. The ground-truth value of lines and words was calculated manually in our database. The statistics of the performance measure for words, as shown in Table III, indicate that the segmentation rate was 99% successful (SR); this segmentation rate is divided into over-segmentation, correct and incorrect segmentation. The accurate segmentation rate (CR) for words is 89%; Recall (Re), precision (P) and F-measure (Fm) obtained approximately 1.00, respectively. The performance measure indicates that the incorrect segmentation rate is approximately 11%; it means that the number of over-segmentations is 985 words (NCo), and 833 words are missed or incorrect segmentation (NCi). As for segmentation, the number of touching Arabic letters is 79404. These letters are either isolated from the word's origin or have single or multiple touching. Table III also indicates that 99% of the total segmented Arabic letters, 87% of the correct segmentation (CSR), and 13% of them was wrong segmentation. 79404 of touching characters were segmented. (P) was 0.96. (Fm) and (Re) were 1.00. Table IV shows the types of challenges in Arabic handwriting obtained from Table III.

Table V shows the comparison results between the outputs of this work with previous works that discussed such problems. On the other hand, the comparison results are difficult because of the different documents collected by researchers from various Arabic handwritten sources that involve these relevant challenges. The primary purpose is to reach satisfactory and accurate solutions even though the collected documents differ, but the essence of the problem is common. It should be noted that the comparison results are based on the findings of the researchers' works.

TABLE IV. SHOWS THE ACCURACY OF THE HYBRID METHOD IN EACH TYPE OF INPUTS

Type of Characters	No of characters	Correct Segmentation	Accuracy
Isolated	28401	26696	94%
Single Touching	32573	28012	86%
Multiple Touching	18430	15186	82%
Total	79404	69452	87.4%

TABLE V. COMPARISON ANALYSIS

Reference	Technique	No. of images	Types of Input	Segmentation ratio
(Aouadi & Kacem, 2017) [22]	Nearest model selection and detect the model's parts centres	820	Segment touching lines and words.	94%
(Ullah et al., 2019) [4]	Overlapping Set Theory and Contour Tracing	220	Segment a single touching character	97.2 %
Proposed method	Hybrid approach	huge	Segment a overlapping and single/ multiple touching characters	90%

VI. CONCLUSION

Characters segmentation in Arabic handwritten is a complex task. However, to improve the efficiency of the recognition system, it is desirable to enhance the character segmentation process in Arabic handwritten. To this extent, this article proposed a hybrid approach for Arabic handwritten character segmentation considering the overlapping or multiple touching issues. The proposed approach aims at improving the efficiency of the segmentation stage. The evaluation is conducted over Arabic handwritten databases which contain more complex challenges than the previously existing ones. The results showed that the proposed approach was highly efficient in word segmentation. And it is an effective, feasible and flexible approach in the segmentation of interconnected, overlapping or multi-touching Arabic characters. But errors or over-segmentation may occur in multi-touched characters. Therefore, in future work, the problem of over-segmentation of multi-touch characters will be considered. Pre-processing operations have also to be improved in formatting very poor documents.

REFERENCES

[1] Qaroush, Aziz, et al. "An efficient, font independent word and character segmentation algorithm for printed Arabic text." Journal of King Saud University-Computer and Information Sciences 34.1 (2022): 1330-1344.

[2] Jindal, Payal, and Balkrishan Jindal. "Line and word segmentation of handwritten text documents written in Gurmukhi script using mid point detection technique." 2015 2nd International Conference on Recent Advances in Engineering & Computational Sciences (RAECS). IEEE, 2015.

[3] Louloudis, Georgios, et al. "Text line and word segmentation of handwritten documents." Pattern recognition 42.12 (2009): 3169-3183.

[4] Naeem-Ullah, Unsar, et al. "First authentic report of Spodoptera frugiperda (JE Smith)(Noctuidae: Lepidoptera) an alien invasive species from Pakistan." Applied Sciences and Business Economics 6.1 (2019): 1-3.

[5] Louloudis, Georgios, et al. "Text line and word segmentation of handwritten documents." Pattern recognition 42.12 (2009): 3169-3183.

[6] Shamsan, Ehab A., et al. "Off line Arabic handwritten character using neural network." 2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA). IEEE, 2017.

[7] Zeki, Ahmed M. "The segmentation problem in arabic character recognition the state of the art." 2005 International Conference on Information and Communication Technologies. IEEE, 2005.

[8] Kang, Le, and David Doermann. "Template based segmentation of touching components in handwritten text lines." 2011 International Conference on Document Analysis and Recognition. IEEE, 2011.

[9] Hamid, A. and R. Haraty. A neuro-heuristic approach for segmenting handwritten Arabic text. in Proceedings ACS/IEEE international conference on computer systems and applications. 2001. IEEE.

[10] Belaïd, A. and N. Ouwayed, Segmentation of Ancient Arabic Documents, in" Guide to OCR for Arabic Scripts," Eds. Volker Märgner and Haikal El Abed. 2011, Springer-Verlag, London.

[11] Farulla, G.A., N. Murru, and R. Rossini, *A fuzzy approach to segment touching characters.* Expert Systems with Applications, 2017. 88: p. 1-13.

[12] Radaideh, A.A. and M.S.M. Rahim, *Existing techniques in Arabic characters recognition (ACR).* Journal of Informatics and Mathematical Sciences, 2016. 8(5): p. 347-360.

[13] Ahmad, I. and G.A. Fink. Class-based contextual modeling for handwritten Arabic text recognition. in 2016 15th international conference on frontiers in handwriting recognition (ICFHR). 2016. IEEE.

[14] Brodic, D. and Z.N. Milivojevic, Text line segmentation with the algorithm based on the oriented anisotropic Gaussian kernel. Journal of Electrical Engineering, 2013. 64(4): p. 238.

[15] Surinta, O., et al. A path planning for line segmentation of handwritten documents. in 2014 14th International Conference on Frontiers in Handwriting Recognition. 2014. IEEE.

[16] Sanchez, A., et al. Text line segmentation in images of handwritten historical documents. in 2008 First Workshops on Image Processing Theory, Tools and Applications. 2008. IEEE.

[17] Hashrin, C., et al. Segmenting Characters from Malayalam Handwritten Documents. in 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT). 2019. IEEE.

[18] Alaei, A., U. Pal, and P. Nagabhushan, *A new scheme for unconstrained handwritten text-line segmentation.* Pattern Recognition, 2011. **44**(4): p. 917-928.

[19] S Deshmukh, Manisha, and Satish R Kolhe. "A hybrid character segmentation approach for cursive unconstrained handwritten historical Modi script documents." Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India. 2019.

[20] Mahmoud, Sabri A., et al. "KHATT: An open Arabic offline handwritten text database." Pattern Recognition 47.3 (2014): 1096-1112.

[21] Pechwitz, M., et al. IFN/ENIT-database of handwritten Arabic words. in Proc. of CIFED. 2002. Citeseer.

[22] Aouadi, N. and A. Kacem, *A proposal for touching component segmentation in Arabic manuscripts.* Pattern Analysis and Applications, 2017. 20(4): p. 1005-1027.