Segmentation of Touching Arabic Characters in Handwritten Documents by Overlapping Set Theory and Contour Tracing

Inam Ullah[1], Mohd Sanusi Azmi[2]
Mohamad Ishak Desa[3]
Faculty of information technology and Communication
Universiti Teknikal Malaysia (UTeM)
Melaka, Malaysia

Yazan M. Alomari[4]
Department of Management Information Systems College of Applied Studies and Community Services
Imam Abdulrahman Bin Faisal University
Dammam, Saudi Arabia

Abstract—**Segmentation of handwritten words into characters is one of the challenging problem in the field of OCR. In presence of touching characters, make this problem more difficult and challenging. There are many obstacles/challenges in segmentation of touching Arabic handwritten text. Although researches are busy in solving the problem of segmentation of these touching characters but still there exist unsolved problems of segmentation of touching offline Arabic handwritten characters. This is due to large variety of characters and their shapes. So in this research, a new method for segmentation of touching Arabic Handwritten character has been developed. The main idea of the proposed method is to segment the touching characters by identifying the touching point by overlapping set theory and ending points of the Arabic word by applying some standard morphology operation methods. After identifying all the points, segmentation method is applied to trace the boundaries of characters to separate these touching characters. Experiments were conducted on touching characters taken from different data sets. The results show the accuracy of the proposed method.**

Keywords—Offline handwritten characters; touching characters; segmentation; overlapping set theory; morphological operation

I. Introduction

Modern age, also called the age of information technology because the computer has importance in every field of life. Computer is considered as essential part of human life. Although it is true that computer have not much intelligent compare to human. Human can recognize any type of text image from historical and degraded documents lying in the libraries while computer can't directly understand these text images [1]. Many efforts are required to convert these historical documents to machine understandable format [2][3] because it is not sufficient to store the information in image format [4]. Researchers are busy in developing new algorithms for segmentation of touching Arabic characters, but still this is long standing problem for conversion of handwritten images to electronic form [5][6]. The problem becomes more serious, when dealing with touching handwritten Arabic words because still there is a gap between human and machine abilities in reading handwriting text under noisy conditions especially for overlapped Arabic manuscripts. This due to the nature of font and style of the Arabic characters, which is written from right to left and is always cursive in both machine printed and handwritten text [7][8][9]. Numerous attempts have been made for the recognition/segmentation of overlapped words in Arabic and other languages as well but these overlapped characters still exist gaps [10]. All these efforts emerge the idea of Optical Character Recognition (OCR).

OCR is a technology that is used to convert paper scanned or other types of images to editable format [11][12]. But before converting these images to editable images it needs image segmentation. Image segmentation is one of the important step in OCR because segmentation subdivides an image and distinguish the area of interest and ignore unwanted information [13]. Although image segmentation is not directly related to recognize the segmented images but they are closely related to each other. Image segmentation is important basis for image recognition [14]. If image segmentation is accurate recognition rate is high otherwise recognition ratio is low.

Segmentation which is used to break the text into lines, words and characters of handwritten text is still a challenging task because handwriting is natural and differs from person to person; therefore, many researchers are investigating solutions to solve the problem and some of them have made remarkable achievements, still more research is needed to improve the performance of already developed systems. To discuss all developed methods in this paper is not possible but research done by, address the issues of touching Arabic handwritten characters.

The rest of the paper is organized as follows. Section II covers basic background about the properties of Arabic language with touching types in the Arabic handwritten documents. Section III describes the related works. Algorithm details are in Section IV. Experimental results are reported in Section V. Conclusion and future work is discussed in Section VI.

II. Properties of Arabic Language

A. Arabic Alphabets

In Arabic language consists of 28 alphabets[15][16] shown in Fig. 1.

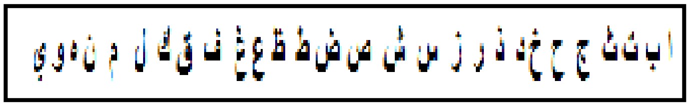


Fig. 1. Arabic Alphabets [15] [16].

B. Shape of Characetrs in Words

Shape of alphabets according to location is shown in Table I. Some letters have the same shape but numbers, location (above and below) of dots and strokes differ one alphabets from other e.g. Ba (ب), Ta (ت), and Tha (ث).

C. Touching Character and its Types

Characters may be joined with the character of other word or with in the same words to form simple touching [18] as shown in Fig. 2.

TABLE I. ARABIC CHARACTER FORMS [17]

Characters	Isolated	Beginning	Middle	End
Alif	ا	-	-	ـا
Ba	ب	بـ	ـبـ	ـب
Ta	ت	تـ	ـتـ	ـت
Tha	ث	ثـ	ـثـ	ـث
Jeem	ج	جـ	ـجـ	ـج
Ha	ح	حـ	ـحـ	ـح
Kha	خ	خـ	ـخـ	ـخ
Dal	د	-	-	ـد
Thal	ذ	-	-	ـذ
Ra	ر	-	-	ـر
Zai	ز	-	-	ـز
Seen	س	سـ	ـسـ	ـس
Sheen	ش	شـ	ـشـ	ـش
Swad	ص	صـ	ـصـ	ـص
Dwad	ض	ضـ	ـضـ	ـض
Tah	ط	طـ	ـطـ	ـط
Dha(Zwa)	ظ	ظـ	ـظـ	ـظ
Ain	ع	عـ	ـعـ	ـع
Ghain	غ	غـ	ـغـ	ـغ
Fa	ف	فـ	ـفـ	ـف
Qaf	ق	قـ	ـقـ	ـق
Kaf	ك	كـ	ـكـ	ـك
Lam	ل	لـ	ـلـ	ـل
Meem	م	مـ	ـمـ	ـم
Noon	ن	نـ	ـنـ	ـن
Ha	ه	هـ	ـهـ	ـه
Waw	و	و	ـو	ـو
Ya	ي	يـ	ـيـ	ـي

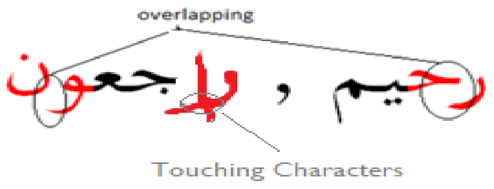


Fig. 2. Touching Character.

Characters can be joined in such a way to form even complicated touching, such as writing in Arabic calligraphy shown in Fig. 3.

Touching of characters taken from handwritten Arabic AHDB dataset is shown in Fig. 4.

Thus, the possible types of touching that normally exists between characters are shown in Table II.

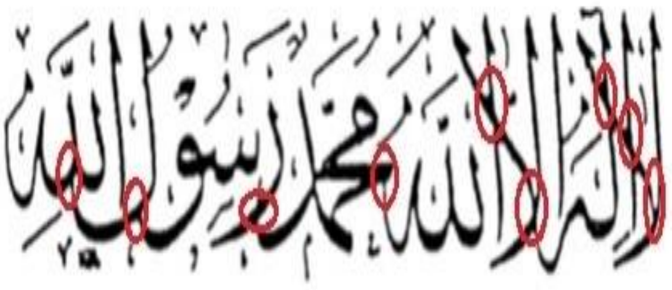


Fig. 3. Complicated Touching Characters.

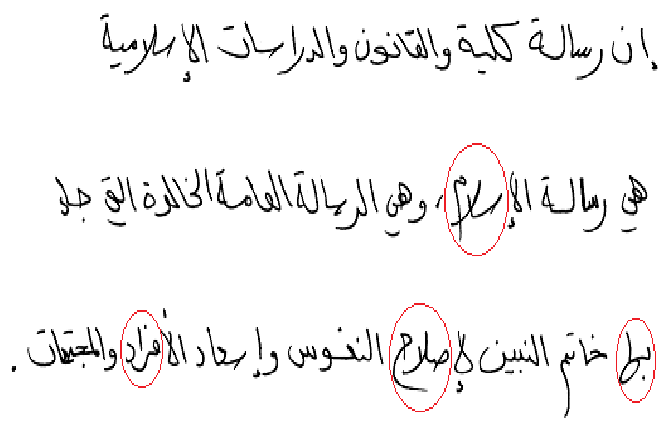
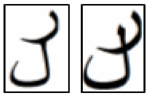
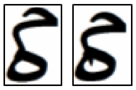
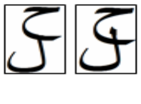
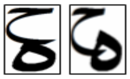


Fig. 4. Touching Characters in AHDB Dataset.

TABLE II. GENERAL TOUCHING TYPES [19]

Type	Letters	Sample
A	Top:[ر, ز, س, ش, ص, ض, ن, ق, و, ى, ي] Bottom: [ا, ط, ظ, ك, ل]	
B	Top: [ر, ز, م, و] Bottom: [ص, ض, ة]	
C	Top: [ج, ح, خ, ع, غ] Bottom: [ا, ط, ظ, ك, ل]	
D	Top: [ج, ح, خ, ع, غ] Bottom: [ه]	

III. Related Work

By exploring the published literature related to segmentation of touching handwritten Arabic characters, there are number of methods proposed for segmentation of handwritten and printed touching text. Some of them are.

In [19], they proposed a new method on the basis of morphological analysis for segmentation of touching lines in Arabic handwritten document. Four types of touching types in Arabic document. Image is converted into skeleton image and the curved is traced by angular variance starting from the starting point. The purpose is to trace the skeleton image in right direction. This method is working for selected touching but not for all touching types because of varieties in handwritten Arabic characters.

In [20], they proposed a new method called template based segmentation method for segmentation of touching characters. In this method created a dictionary file, which contains template of all possible touching with necessary detail. Then compare with input image and select template from the dictionary file. This method works well some image but can’t segment all touching characters.

Problems identified in this method:

1) Tedious method because of creation of template file.

2) Template dependent method, will not segment all touching characters.

3) Time consuming, because for simple touching characters will search whole dictionary file.

In [21], proposed a new segmentation method for segmentation of touching handwritten digits. First finding the end-points by applying some standard method. Now by tracing the boundary from end-point collecting all co-ordinates and segmentation point. Thus segmented the touching digits with this recognition ratio 71.43%. But problems identified in the proposed method are given below:

1) Proposed method is only for segmentation of numeric digit not used for character data.

2) Proposed method is only for simple touching for multiple touching occurs over-segmentation.

In [22], proposed a new segmentation technique for solving the problem touching handwritten Arabic characters, touching is between two characters in same word or other line. This method is template based and these touching images are compare with the template file already create before. Problems in the proposed method:

1) Performance of proposed method depended on template.

2) As in handwritten text there is a lot of variation even between same writer. Any incorrect template selection affects the recognition rate.

3) Sometime produce broken characters during segmentation.

In [23], proposed method is basically for segmentation and recognition of Arabic touching characters. The concept is to trace the boundary of character. Normalize center of gravity of region. Then calculate minimum distance and apply horizontal distance between connected region. Segment the connected characters. Problem identified:

1) Segmentation of ligature is not explained neither give the results or comparison to clear the methodology.

2) Authors claimed the small modification can be used for segmentation of various documents including Othman script of Al-Quran. But Quran of Othman Script have multiple touching points and also have some broken characters are possible, especially in handwritten script. No solution is given for merging of broken characters

In [14], proposed method for segmentation of touching handwritten and printed Latin characters. Basically this method is a combination of three already developed methods taken from the literature. During database selection for testing, highlighted the problem of unavailability of standard datasets. Although they claim high success segmentation ratio but still this method is only applicable to certain situation or document, which is the drawback of the proposed method.

IV. Proposed Method

In this section, proposed method for segmentation of touching handwritten Arabic character with mathematical background is explained. The model of the proposed method is illustrated in Fig. 5.

In the first stage of the model, input image is converted to binary image and also to enhance the quality of the image by removing unwanted information.

```
Step 1.   Find all endpoints of touching image. In Fig. 6
          Four endpoints in the image $E_1$, $E_{2,}$ $E_3$ and $E_4$.

Step 2.   Find coordinate between any two Endpoints by
          tracing boundary of image. Suppose between $E_1$ ,
          $E_2$ and $E_3$, $E_4$. Thus
           Set A = {Coordinates between $E_1$ , $E_2$ }
           Set B = {Coordinates between $E_3$ , $E_4$ }

Step 3.   Apply overlapping set theory on  Set A and B
           Set C = Set A ∩ Set B
            If  IsEmpy Set C
            No touching character
          Else
            Touching character and element of Set C is the
            touching point.

Step 4.    END
```

In the second stage of the model, End-points, touching and neighboring point are detected in the input image. For endpoints and neighboring points of thinned image applied standard method but for touching point applied overlapping set theory. Steps of overlapping set theory are to find the junction or touching point.

After finding the touching point, next is the neighbor point in all direction of the image near touching point. Touching and neighboring points are shown in Fig. 6 and Fig. 7 below.

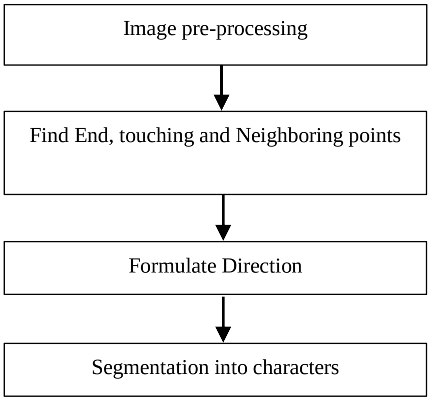


Image pre-processing
Find End, touching and Neighboring points
Formulate Direction
Segmentation into characters

Fig. 5. Proposed Method Model.

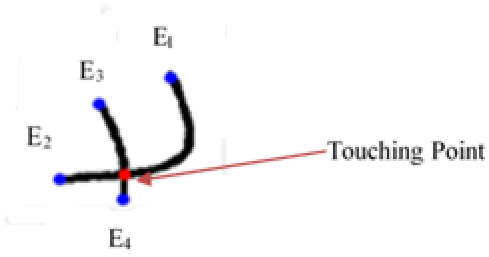


Fig. 6. Endpoints and Touching Point.

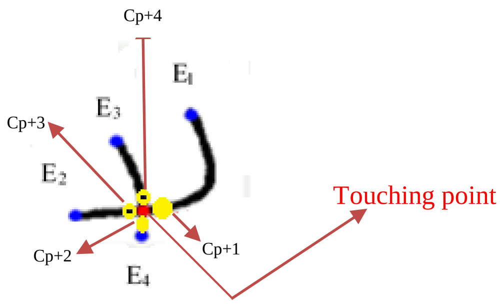


Fig. 7. Neighboring Points.

In the third stage (Formulate direction), correct boundary of touching character is identified by neighbor points. The purpose of this step is to trace the boundary of touching character in correct direction because at the location of touching point there are many paths or curves. The coordinates of neighboring points C_{p+1} , C_{p+2} , C_{p+3} and C_{p+4} help in selection of right direction. Fig. 8 shows the selection of curve near touching point.

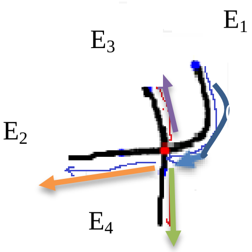


Fig. 8. Formulate the Direction.

The proposed method is based on contour tracing starting from one endpoint and continue tracing the boundary of character to the touching point(C_p). At this point, there are three possible directions (see Fig. 8) towards endpoint E_2 , E_3 and E_4. Here the neighbor points play an important role for direction. Therefore tracing character boundary from E_1 will follow the curve towards endpoint E_2 because by comparing the coordinates of neighboring point C_{p+1} with other neighboring points C_{p+2} , C_{p+3} and C_{p+4} , there is a abrupt change towards point C_{p+2} and C_{p+4}. While normal change towards neighboring point C_{p+3}. Fig. 8 shows three curves near touching point.

At this stage, endpoints, touching point, neighboring points and also curve direction of the touching image are identified. Next stage is segmentation of touching character image into separate characters. Fig. 9 explains the segmentation algorithm flowchart.

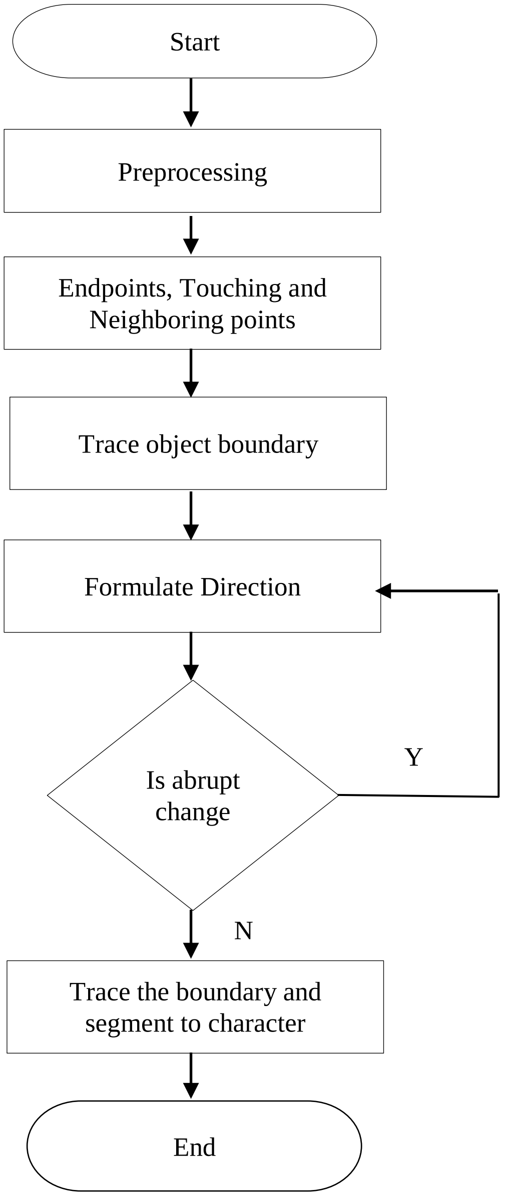

Start
Preprocessing
Endpoints, Touching and Neighboring points
Trace object boundary
Formulate Direction
Is abrupt change
Y
N
Trace the boundary and segment to character
End

Fig. 9. Flowchart.

The Explanation of the flowchart in Fig. 9 can be detailed as follow:

```
Algorithm : Segmentation of touching
  characters

Start
Step 1:    Input an image
  Image is handwritten or printed
  touching or without touching
  characters.

Step 2: Perform Pre-processing
  Very important step to enhance the
  quality of image and remove
  unwanted information.

Step 3: Find End-Points and Touching
  point
  Using overlapping set theory to
  find the touching.

Step 4: Find Neighboring points near
  Touching point
  Here find neighbor points in all
  direction, near the junction or
  common point

Step 5: Formulate the direction
  The purpose of this step is to
  trace the boundary of touching
  character in correct direction
  because at the location of
  touching point there are many
  paths or curves.

Step 6: Segmentation of the touching
  characters
  At this step touching character
  image is segmented to separate
  characters.

End
```

V. Results and Discussion

Basically, this proposed method is for segmentation of touching handwritten Arabic characters. As due to lack of standard dataset for handwritten data. Touching handwritten Arabic data are collected from different datasets, while some are converting to touching by doing some manually work. In section A list of datasets for testing the proposed method. Section B is for comparison with others two selected methods, which are closely related to segmentation of touching characters.

A. *List of Dataset for this Research*

Data collection is very important to test the performance of the proposed method. Due to unavailability of standard touching character datasets. Selected number of dataset, details of the collected data are shown in Table III below.

TABLE III. Data Collection

Database	Purpose
AHDB	Off-line handwriting
IFN/ENIT	Tunisian city names Off-line handwriting
Arabic handwritten 1.0	Off-line handwriting
IBN SINA	Arabic Manuscript
IAM	English Handwritten. Total number of writers, 657 collected handwriting samples. Number of isolated and labelled word 115320.
NIST	There are 150,000 handwritten binary digits number of broken digits 2600

Touching character were selected manually from the datasets in Table III, especially from AHDB as shown in Fig. 4. While in some cases did some manual work in order to get some challenging types of two touching characters to test the performance of the proposed method.

B. *Analysis*

In this section presented results of the experiments, which were conducted to prove the performance of the proposed method. Comparison of proposed with that of other methods are shown in Table IV.

It shows from the experiments that proposed method is very flexible and efficient for segmentation of touching Arabic handwritten characters. Samples results are given in Fig. 10 below.

TABLE IV. Analysis

Touching Characters	Proposed Method		
	Total Selected Images	Segmented Images	Segmentation Percentage
Other Method 1 [22] inter-word touching.	220	--	94%
Other Method 2 [19] inter-word touching.	622	620	96.88%
Proposed method	220	214	97.27%

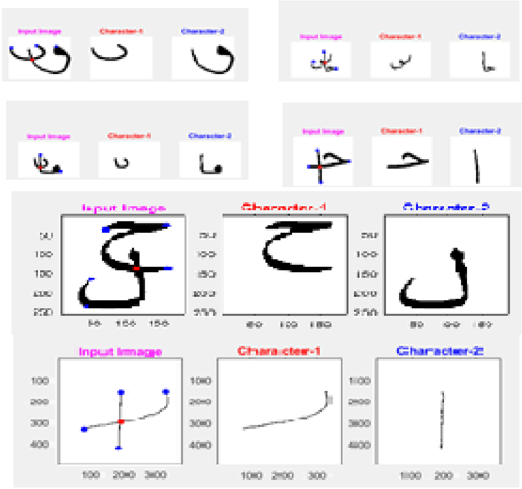


Fig. 10. Samples of Results.

Based on Fig. 10, the proposed method is correctly segmented touching handwritten Arabic character in an efficient way. Total number of images selected 220 of almost every type of touching character and out of these 220 images correctly segmented 214 images. Only 6 images (2.73%) either over or under segmented because of varieties in Arabic handwritten data.

VI. Conclusion

Touching handwritten character images normally exist in old historical documents or writing text in Arabic calligraphy. These touching characters extensively happen in English, numbering and Arabic handwritten historical materials. However, for this research will considered only Arabic handwritten characters and numeric data. Thus proposed method is only for segmentation of touching Arabic handwritten characters to solve the longstanding and unsolved problem of segmentation of touching characters. In final conclusion, to say that proposed method is very flexible and if improved further can be used for multiple touching Arabic characters while in future work can be used for other languages that are similar to Arabic such as Urdu, Pashto and Farsi languages.

Acknowledgment

The authors would like to thank Ministry of Education for funding this study through the following grant FRGS/1/2017/ICT02/FTMK-CACT/F00345. We would also like to extend our thanks to Universiti Teknikal Malaysia Melaka and Faculty of Information Technology and Communication for providing excellent research facilities.

References

[1] S. A. Malik, M. Maqsood, and F. Aadil, Advances in Information and Communication, vol. 70. Springer International Publishing, 2020.

[2] H. Modi and M. C., “A Review on Optical Character Recognition Techniques,” Int. J. Comput. Appl., vol. 160, no. 6, pp. 20–24, 2017.

[3] C. S. Lwin and W. Xiangqian, “Image Purification Technique for Myanmar OCR Applying Skew Angle Detection and Free Skew,” Int. J. Sci. Res. Sci. Technol., vol. 6, no. 1, pp. 186–203, 2019.

[4] N. Aouadi, S. Amiri, and A. K. Echi, “Segmentation of Connected Components in Arabic Handwritten Documents,” Procedia Technol., vol. 10, pp. 738–746, 2014.

[5] T. Saba and A. Rehman, “Character Segmentation in Overlapped Script using Benchmark Database,” pp. 140–143.

[6] A. Gattal, Y. Chibani, and B. Hadjadji, “Segmentation and recognition system for unknown-length handwritten digit strings,” Pattern Anal. Appl., vol. 20, no. 2, pp. 307–323, 2017.

[7] J. H. Alkhateeb, “A Database for Arabic Handwritten Character Recognition,” Procedia Comput. Sci., vol. 65, no. Iccmit, pp. 556–561, 2015.

[8] A. Amin, “Off-line Arabic character recognition,” Pattern Recognit., vol. 31, no. 5, pp. 517–530, 2002.

[9] S. Ahmed, S. Naz, M. Razzak, and R. Yusof, “Arabic Cursive Text Recognition from Natural Scene Images,” Appl. Sci., vol. 9, no. 2, p. 236, 2019.

[10] M. S. Deshmukh and S. R. Kolhe, “A Hybrid Character Segmentation Approach for Cursive Unconstrained Handwritten Historical Modi Script Documents,” pp. 967–978, 2019.

[11] N. Vincent and J. M. Ogier, “Shall deep learning be the mandatory future of document analysis problems?,” Pattern Recognit., vol. 86, pp. 281–289, 2019.

[12] M. Ayesh, K. Mohammad, A. Qaroush, S. Agaian, and M. Washha, “A Robust Line Segmentation Algorithm for Arabic Printed Text with Diacritics,” Electron. Imaging, vol. 2017, no. 13, pp. 42–47, 2017.

[13] S. Eskenazi, P. Gomez-Krämer, and J. M. Ogier, “A comprehensive survey of mostly textual document segmentation algorithms since 2008,” Pattern Recognit., vol. 64, no. October 2016, pp. 1–14, 2017.

[14] G. A. Farulla, N. Murru, and R. Rossini, “A fuzzy approach to segment touching characters,” Expert Syst. Appl., vol. 88, pp. 1–13, 2017.

[15] S. Khan, H. Ali, Z. Ullah, N. Minallah, S. Maqsood, and A. Hafeez, “KNN and ANN-based Recognition of Handwritten Pashto Letters using Zoning Features,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 10, 2018.

[16] S. Wshah, Z. Shi, and V. Govindaraju, “Segmentation of Arabic handwriting based on both contour and skeleton segmentation,” Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, no. January, pp. 793–797, 2009.

[17] Y. M. Alginahi, “A survey on Arabic character segmentation,” Int. J. Doc. Anal. Recognit., vol. 16, no. 2, pp. 105–126, 2013.

[18] K. Anwar, Adiwijaya, and H. Nugroho, “A segmentation scheme of Arabic words with harakat,” 4th IEEE Conf. Commun. Networks Satell. COMNESTAT 2015 - Proc., pp. 111–114, 2016.

[19] N. Ouwayed and A. Belaïd, “Separation of overlapping and touching lines within handwritten arabic documents,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 5702 LNCS, pp. 237–244, 2009.

[20] L. Kang and D. Doermann, “Template based segmentation of touching components in handwritten text lines,” Proc. Int. Conf. Doc. Anal. Recognition, ICDAR, pp. 569–573, 2011.

[21] A. N. G. Lopes Filho and C. A. B. Mello, “Segmentation of Overlapping Digits through the Emulation of a Hypothetical Ball and Physical Forces,” pp. 223–226, 2015.

[22] N. Aouadi and A. Kacem, “A proposal for touching component segmentation in Arabic manuscripts,” Pattern Anal. Appl., pp. 1–23, 2016.

[23] Z. Saber, A. Q. Sabri, A. Kamsin, and S. Hakak, “Efficient Approach to Segment Ligatures and Open Characters in offline Arabic Text,” Int. J. Comput. Commun. Instrum. Eng., vol. 4, no. 1, pp. 40–44, 2017.