Junction Point Detection and Identification of Broken Character in Touching Arabic Handwritten Text using Overlapping Set Theory

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

***Abstract*—Touching characters are formed when two or more characters share the same space with each other. Therefore, segmentation of these touching character is very challenging research topic especially for handwritten Arabic degraded documents. This is one of the key issue in recognition of the handwritten Arabic text. In order to make the recognition system more effective segmentation of these touching handwritten Arabic characters is considered to be very important research area. In this research, a new method is proposed, which is used to identify the junction or common point of Arabic touching word image by applying overlapping or intersection set theory operation, which will help to trace the correct boundary of the touching characters, identify the broken characters and also segmented these touching handwritten text in an efficient way. The proposed method has been evaluated on Arabic touching handwritten characters taken from handwritten datasets. The results show the efficiency of the proposed method. The proposed method is applicable to both degraded handwritten documents and printed documents.**

Keywords—Touching characters; segmentation and recognition; overlapping set theory; junction point; broken character

I. Introduction

Recognition of handwritten characters, which is challenging problem in the field of pattern recognition [1]. Mostly characters segmentation techniques are used to recognize these segmented text and it is considered very important step for recognition because incorrect segmentation effects the recognition [2][3]. Almost everywhere in the world, Libraries and National archives have huge volume of historical and degraded documents in the form of books. These valuable materials need proper attention in conversion to machine readable format [4]. Although, researchers are busy in solving the problems of these documents in order to provide them in meaningful form for further studies, researches and projects. Some of the work done in this regard is only by scanning of these documents, which is not sufficient to store these information in image format [5]. It requires more research work to convert these historical handwritten and degraded documents into machine understandable format. Therefore, it is still considering as an open and important research area. The big problem facing by researchers is by not following the standard rules of writing in these documents especially in Arabic handwritten documents. Which are facing a lot of difficulties because of nature and style of Arabic language characters, where characters are connected and direction of writing is from right to left. There is no upper or lowercase letters both in printed and handwritten Arabic writing [6][7].

Arabic language consists of 28 characters and every individual character has a fixed shape. As, in Arabic writing characters are connected with each other to form words, these connections change the shape of the characters i.e. shape of isolated character is different than character in middle and end of the word [8]. In Arabic handwriting, normally the writers don't follow the standard writing rules, means the writer is free to write according to his well, situation etc. so there are great possibilities in writing a document that characters may touch, overlap and not properly written and produce broken character. So, in presence of these problems conversion of handwritten text to electronic form is not fully succeeded and problem becomes more serious, when dealing with touching Arabic handwritten words because still there is a gap between human and machine abilities in reading handwriting text under noisy conditions especially for touching and broken Arabic manuscripts. These challenging situation i.e. Connected/ touching components complicate the segmentation and recognition process because of unavailability of databases [9] and also some characters are combine with each other in such a way that it forms ligatures [10].

The rest of the paper is organized as follows. Section II covers possible touching character 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. Touching Characters and its Types

As Arabic writing, characters are connected with each other. It creates a challenges situation by comparing with other languages such as English. Therefore, general writing style [11] and some of the touching and overlapping of characters are shown in Fig. 1, Fig. 2 and Fig. 3 [12].

In Arabic handwritten documents, there are four overlapping/touching types as shown in Table I.

Some of the touching characters are highlighted in Fig. 4 [14].

Fig. 1. Overlapping of Characters with Touching.

Fig. 2. Overlapping without Touching.

Fig. 3. Touching by Mistake.

Based on Fig. 4 and Table I, some of the possible touching structures are shown in Fig. 5.

TABLE. I. TOUCHING TYPES IN ARABIC HANDWRITING [13]

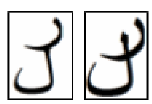
Type	Letters	Sample
A	Top:[ر, ز, س, ش, ص, ض, ن, ق, و, ى, ي] Bottom: [ا, ط, ظ, ك, ل]	
B	Top: [ر, ز, م, و] Bottom: [ص, ض, ة]	
C	Top: [ج, ح, خ, ع, غ] Bottom: [ا, ط, ظ, ك, ل]	
D	Top: [ج, ح, خ, ع, غ] Bottom: [ه]	

Fig. 4. Touching Character in Arabic Handwriting.

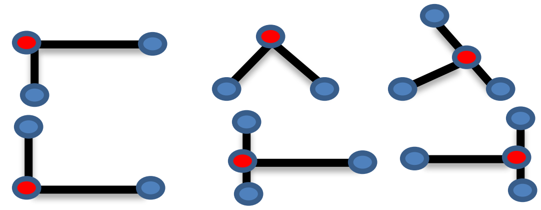


Fig. 5. Possible Touching Structure.

Where represents end-points, Junction point and Boundary of character.

III. Related Work

By exploring the published literature related to junction point detection, there is a lack of research work for handwritten and printed Arabic characters but number of methods proposed for other languages such as Chinese, English and junction point detection for general images. Some of the methods are discussed below.

In [15] proposed method is for stroke extraction of handwritten Chinese characters. In proposed method first of all character image is converted to binary image. Then apply thinning algorithm to converted binary image. Next is to find endpoints and fork point of the skeleton image and then these points are checked with the original image of handwritten Chinese character. Main objectives of this proposed method is to increased segmentation and recognition ratio.

In [16], proposed a junction detection method, which is used for solving the problem of segmentation of touching characters i.e. two touching character string. This proposed method is used only for uppercase printed English Alphabets and this method is applicable to only single touching not working for multiple touching.

Some of the limitations of this proposed method are:

1) Segmentation of English Uppercase character.
2) Used for selected printed characters.
3) Used for segmentation of single touching.
4) Not working for multiple touching and handwritten characters.

In [17] Junction based approach for solving the problem of touching remote sensing images. New formula also include minimization criteria for the total weighted distance is proposed. This will detect junction point accurately. Authors claim that its results are much better than popular junction detection detector such as Forstner, JUDOCA (Junction Detection Operator based on circumferential Anchor) and CPDA.

In [18] Junction based approach for solving the problem of touching Chinese characters. The main drawback of this method it is only used for identification not segmentation. In Arabic language characters are connected with each other's thus form sub words in the same word. Here in this work i.e. identification of junction point only collect information around this point and ignore the whole shape of a character. This is very important step in recognition or writer identification.

In [19] Junction based approach for solving the problem of touching Handwritten Devanagari Character Recognition. Finding the junction point where chain code moves in more than one direction in 8-connected neighborhood or in other word a point which have more than two neighboring pixel. The main drawback with this method is can't apply to Arabic language because characters are connected with each for form word or piece of word.

In [20] authors have proposed an improved skeletonization method. This method is basically the combination of straight and cures lines. Its main purpose to improve the junction point detection and topological error near the junction point. This method is only applied for architectural drawings and not for handwritten character images.

IV. Proposed Method

In this section is discussed the proposed method of finding the junction point and identification of broken character in touching handwritten Arabic character images [21]. The outline of the proposed method is shown in Fig. 6.

Stage-1: First stage of the model is Image preprocessing. As the proposed method is working on binary images. Thus first step is to convert the input image into binary by applying standard Otsu's method. Binary image is that type of image in which each pixel has two possible values 0 and 1, in other word, this image has two colors background and foreground. Next step is to apply thinning algorithm This Thinning operation is also called skeletonization. This is a morphological operation used to remove selected foreground pixels from binary image. Its main objective is to preserve the important information and make it easy for further processing. Therefore, the output of thinning process is a binary image of only one-pixel width lines. Preprocessing step is shown in Fig. 7.

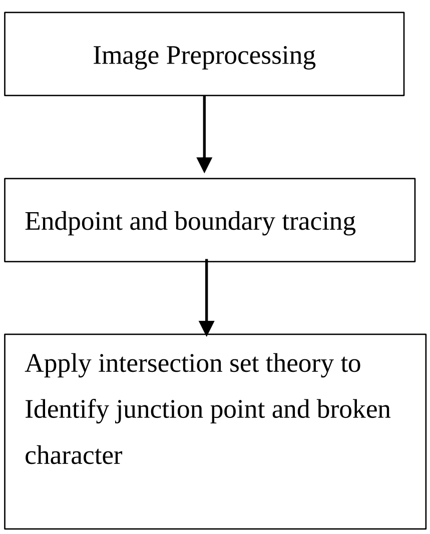


Fig. 6. Proposed Method Stages.

Stage-2: In this stage, Endpoints are those points which have only one neighboring point and which is located at the end of character boundary. In Fig. 8 have touching character image. Fig. 8(a) have four endpoints E_1, E_2 ,E_3 and E_4. While tracing boundary between endpoints gets two sets E_1E_2 and E_3E_4.

Set A $E_1E_2 = \{(x1,y1),(x2,y2)\ldots\ldots(xn,yn)\}$

Set B $E_2E_4 = \{(x1,y1),(x2,y2)\ldots\ldots(xn,yn)\}$

In Fig. 8(b) there are six endpoints E_1, E_2 ,E_3, E_4 , E_5 , E_6 and three sets E_1E_3, E_2E_5, E_4E_6.

Fig. 7. Binary and Thinned Image.

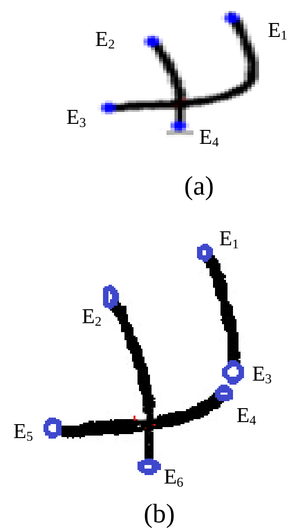


Fig. 8. Touching Character Image.

Set A $\mathbf{E_1E_3 = \{(x1,y1),(x2,y2)\ldots\ldots(xn,yn)\}}$

Set B $\mathbf{E_2E_5 = \{(x1,y1),(x2,y2)\ldots\ldots(xn,yn)\}}$

Set C $\mathbf{E_4E_6 = \{(x1,y1),(x2,y2)\ldots\ldots(xn,yn)\}}$

Stage-3: This stage is to apply intersection set theory on sets in Stage-2. By applying intersection set theory on Set A and Set B shown in Fig. 9.

In Fig. 10 both sets have a common element and that is common or touching point. Apply set theory for Set A, Set B and Set C

While in Fig. 11, Set B and Set C have common element that is the Junction point. Set A has no common element that identify that character is broken.

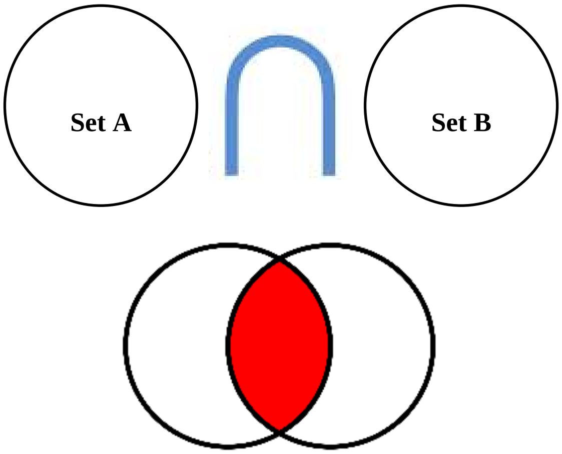


Fig. 9. Overlapping Set Theory.

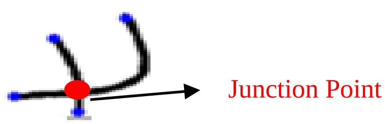


Fig. 10. Junction Point.

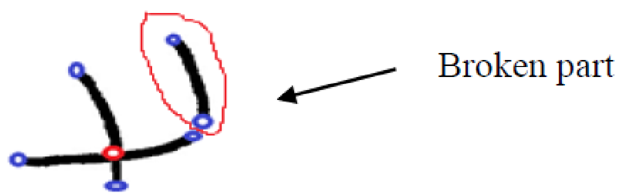


Fig. 11. Identification of Broken Character.

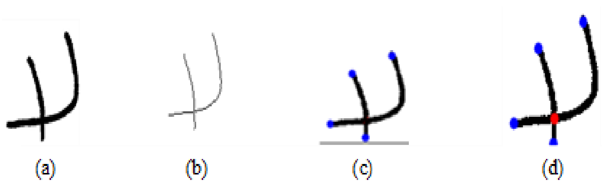


Fig. 12. Output Touching Image with end and Touching Point.

Fig. 12(a), (b), (c) and (d) shows input image, Thinned Image, Endpoints and Junction point.

V. Results and Discussion

The proposed method is for finding junction point and to identify the broken characters but due to lack of research and standardized datasets for testing proposed method, especially for touching characters. Data are collected from different available dataset are shown in Table II and did some manual work to make the touching types more complex shown in Fig. 13.

Total number of sample collected 360 touching Arabic handwritten words according to four different type of touching characters shown in Table I. For each type equal numbers of 90 touching characters were selected. Proposed methods were tested on collected data. In Table III describes the result of each touching type.

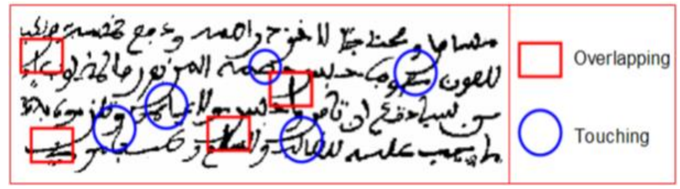


Fig. 13. Manual Touching Collected Data Sample.

TABLE. II. Collected Datasets

Dataset	Size	Purpose
IFN/ENIT	26459 Tunisian City Name	Offline handwritten Text
CEDAR Arabic Dataset	100 pages of text, each comprises 150-200 words	Offline handwritten Text
AHDB		Offline handwritten Text
Arabic-Handwritten 1.0	5000 handwritten pages	Offline handwritten Text

TABLE. III. Results of Finding Common point

Touching Type	Sample	Correctly Identify Common point	Percentage (%)
A	90	88	97.7%
B	90	86	95.55%
C	90	80	88.8%
D	90	82	91.11%

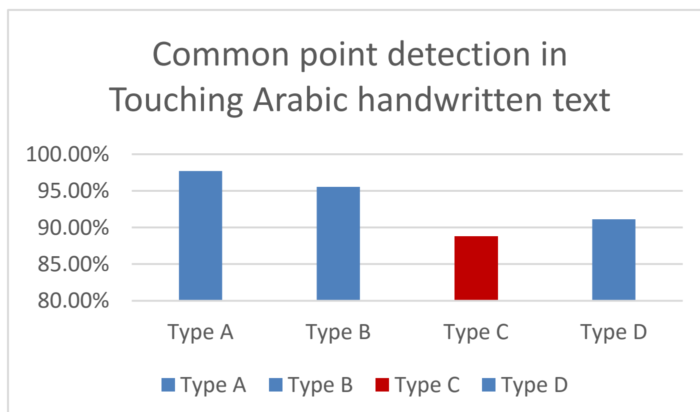


Fig. 14. Analysis of Method.

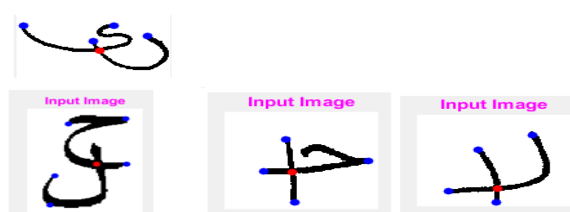


Fig. 15. Sample Output.

The weighted mean of these results is equal to 93.3%, 6.6% rate error because of variety and complexity of Arabic Language. Fig. 14 shows graph of the Table III.

Some of the sample results are given in Fig. 15.

VI. Conclusion

Broken and touching of character images of handwritten exists especially in ancient documents due to the quality of scanning. Thus by exploring the literature for our research we found that broken and touching characters extensively happen in English, Number and Arabic handwritten historical materials. Thus to identify junction point and broken character, proposed new method for finding junction point between the touching characters. This method is based on handwritten Arabic calligraphy where touching of character is most likely occurred. The proposed method is used intersection set theory. This method is not only useful in junction point but also identify the broken character as well. In future this can be applied to others regional languages Urdu, Farsi and many others, which are similar to Arabic language and this method can also be applied to multiple touching characters

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

[2] A. Yamamah and D. Branch, “Cursive Multilingual Characters Recognition Based on Hard Geometric Features Amjad Rehman 1 2,” 2012.

[3] S. Zhao, Z. Chi, P. Shi, and H. Yan, “Two-stage segmentation of unconstrained handwritten Chinese characters,” Pattern Recognit., vol. 36, no. 1, pp. 145–156, 2003.

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

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

[6] A. P. Giotis, G. Sfikas, B. Gatos, and C. Nikou, “A survey of document image word spotting techniques,” Pattern Recognit., vol. 68, pp. 310–332, 2017.

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

[8] F. Khan, A. Bouridane, F. Khelifi, R. Almotaeryi, and S. Almaadeed, “Efficient segmentation of sub-words within handwritten arabic words,” Proc. - 2014 Int. Conf. Control. Decis. Inf. Technol. CoDIT 2014, no. i, pp. 684–689, 2014.

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

[10] Y. Osman, “Segmentation algorithm for Arabic handwritten text based on contour analysis,” Proc. - 2013 Int. Conf. Comput. Electr. Electron. Eng. ’Research Makes a Differ. ICCEEE 2013, pp. 447–452, 2013.

[11] T. Sari, L. Souici, and M. Sellami, “Off-line handwritten Arabic character segmentation algorithm: ACSA,” Proc. - Int. Work. Front. Handwrit. Recognition, IWFHR, no. May 2014, pp. 452–457, 2002.

[12] A. Lawgali, “A Survey on Arabic Character Recognition,” Int. J. Signal Process. Image Process. Pattern Recognit., vol. 8, no. 2, pp. 401–426, 2015.

[13] 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.

[14] B. Al-Badr and S. A. Mahmoud, “Survey and bibliography of Arabic optical text recognition,” Signal Processing, vol. 41, no. 1, pp. 49–77, 1995.

[15] K. Liu, Y. S. Huang, and C. Y. Suen, “Identification of fork points on the skeletons of handwritten chinese characters,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, no. 10, pp. 1095–1100, 1999.

[16] U. K. S. Jayarathna, “A Junction Based Segmentation Algorithm for Offline Handwritten Connected Character Segmentation,” 2006.

[17] J. Zhang, T. Luo, G. Gao, and L. Lian, “Junction Point Detection Algorithm for SAR Image,” Int. J. Antennas Propag., vol. 2013, no. 3, pp. 1–9, 2013.

[18] S. He, M. Wiering, and L. Schomaker, “Junction detection in handwritten documents and its application to writer identification,” Pattern Recognit., vol. 48, no. 12, pp. 4036–4048, 2015.

[19] Arora, S. et al. (2008) ‘Combining multiple feature extraction techniques for Handwritten Devnagari Character recognition’, IEEE Region 10 Colloquium and 3rd International Conference on Industrial and Information Systems, ICIIS 2008. doi: 10.1109/ICIINFS.2008.4798415

[20] Hilaire, Xavier & Tombre, Karl. (2001). Improving the Accuracy of Skeleton-Based Vectorization. 273-288. 10.1007/3-540-45868-9_24.

[21] Inam Ullah and Azmi, M. S. (2019) ‘Segmentation of Touching Arabic Characters in Handwritten Documents by Overlapping Set Theory and Contour Tracing’, (IJACSA) International Journal of Advanced Computer Science and Applications, 10(5), pp. 155–160.