International Journal of Computer Applications Technology and Research
Volume 3– Issue 2, 88 - 92, 2014, ISSN: 2319–8656

A New Approach to Segmentation of On-Line Persian Cursive Words

Golestani, M.R
Department of Computer Engineering, Islamic Azad University South Tehran Branch
Tehran, Iran

Khademi, M.
Department of Applied Mathematics, Islamic Azad University South Tehran Branch
Tehran, Iran

Moeini, A.
Division of Algorithms and Computation, Department of Engineering Science,
College of Engineering,
University of Tehran
Tehran, Iran

Abstract: Segmentation approaches, as processes that divide word into smaller parts which contain one letter at most, have important effect on cursive word recognition. While online cursive word recognition became applied technology in Latin and Chinese languages, complex structural features in Arabic-based script made it an important field of study in Persian and Arabic languages. In this paper, by introducing of Standard Persian Handwriting, we proposed a novel approach to segmentation online Persian cursive script based on width of letter's body in Persian language. Results are shown 99.86% accuracy in detection of expected segmentation points, while recognized extra points reduced 93.73% compared to our previous methods.

Keywords: Segmentation, Persian cursive handwriting, Letters width, Words feature, Online recognition

1. INTRODUCTION

In recent decade, by growth in usage of Pen-Based devices such as Smart phones, Tablets and etc., online handwriting recognition get lots of attentions. Online cursive word recognition studies gain more success in Latin and Chinese languages due to frequent researches [1, 2, 3], while complex structural features in Arabic-based script made online Persian cursive recognition an open field of study.

In Arabic and Persian, words are cursive by nature and every letter could have different shapes based on position of the letter in a word. According to these features and some other like different position of diacritic marks on letters with common body, recognition of Persian and Arabic cursive words are difficult. In recent years, researchers tried to reduce this complexity [4-12].

For reducing complexity, we could break up whole word recognition to smaller parts. Segmentation is a way to divide cursive word into smaller parts which contain one letter at most. By usage of segmentation, one could recognize whole word by combination of small parts recognition result. So segmentation method has important effect on whole word recognition. Recently, different efforts made to propose segmentation method in online Persian and Arabic cursive word recognition [4-11, 13].

In this paper, by introducing of Standard Persian Handwriting, a new approach has been proposed for detecting the segmentation points. The rest of the paper is organized as follows: In next section, we will have a review on researches have been made upon the subject. In Section 3, we introduce Standard Persian Handwriting. In Section 4, we introduce our segmentation method. Section 5, illustrates the experimental results and finally, the paper will be concluded in section 6.

2. LITERATURE SURVEY

Due to effect of segmentation on whole cursive word recognition in Persian and Arabic language, various segmentation methods were presented to reduce recognition complexity by researchers. A list of Persian language features was introduced and used to find segmentation points [6]. By usage of writing direction, points were considered that follow a specific pattern in their direction before and after themselves, as a segmentation point. Series of directions is another feature that was used [6, 7]. The last repeating points in a series of adjutant points, first and last points of each strokes of a word, point's first and second derivatives of right and left, are other features that were used to detect segmentation points in [6].

In [8], according to curve structures in Persian language, curves and their features was detected in input points, then segmentation points defined based on them. Samimi et. all [7] convert input points to their proposed patterns. They defined 7 basic shape including Semicircle in 4 direction, horizontal line, vertical line and oblique line. After conversion of input points to related pattern, segmentation points were detected based on shapes.

Khaled et. all [10] studied on Arabic cursive word. They first defined joint line by angle between line of adjacent points and the horizontal axis and only consider semi-horizontal lines moving from right to left. Then checked above and below of the joints and keep joints that have not any point in above and below. Finally, by integrating these joint lines, middle points were considered as segmentation points.

Two specific features in Persian words were defined in our previous proposed method [13]. In joining of two letters, joining position always be placed over the right-to-left writing direction. Also there is an obvious difference in gradient of ending points of the preceding and beginning points of the succeeding letter. With usage of these features, they first found points with specific gradient on right-to-left direction, then track this points to detect beginning point of next letter and consider it as segmentation point. While segmentation points were detected in this method, some other extra points were detected too.

3. STANDARD HANDWRITING

Persian alphabet contains 32 letters, which has 4 more letters than Arabic. Letters in Persian and Arabic languages could have different shapes based on position of the letter in a word.

Every letter could have utmost four different shapes for isolated, initial, middle and final modes. All different shapes of all Persian letters illustrated in [12].

In addition, various and different writing styles in Persian handwriting, made Persian handwriting recognition complicated. In some samples people write letters totally different from letter main structure or in some other they deform a letter and turn it to some other letters. For example they write letter "د" same as letter "ر ". In some other samples letters were written inside each other and have overlap.

To prevent this undesirable complexity, we introduce a Standard Persian Handwriting that will be applied on handwriting samples. A standard Persian Handwriting must include 3 below condition:

- Letter shape must be written same as normal Persian letter shape. For example letter "د" must not write same as letter "ر".

- Joining letters must have not overlap.

- Words must be written on a straight horizontal line. It means that each part of word did not write in different position. Parts of word must be written along each other same as normal Persian language.

So, if someone consider normal Persian language and do not create a new shape for letters, handwriting will be Standard Persian Handwriting. In this research, all writers followed Standard model and proposed method is based on Standard Persian Handwriting samples.

4. PROPOSED ALGORITHM

In this Section, by introducing a special feature in Persian language, we present a new method for detection of segmentation points in Persian cursive words. In the following, we first depict the feature, and then present our algorithm.

4.1 The Feature

Width of letters body in Persian language has difference in different parts of the letter. If we just consider beginning part of letter, usually have bigger width than situation that we just consider end part of letter. Figure 1 shows different widths in various parts of a word. In joining of two letters, joining position has lowest width as shown in Figure 1. This feature was confirmed for all Persian letters joining. Joining position is best candidate for segmentation points because segmentation process goal is to divide words to letters or basic shapes.

According to this feature, one could detect all positions that letters joint together in a word. Although some other parts of words could include same feature, too.

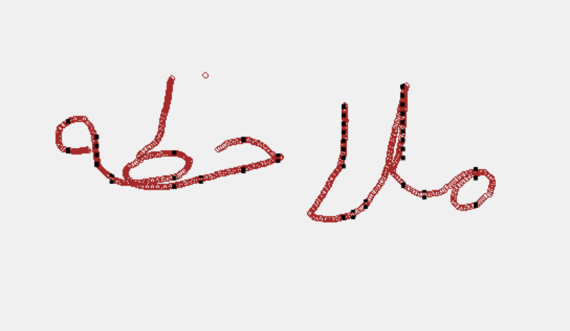

Figure. 1 Different width in various parts of word "ملاحظه"

4.2 Segmentation Algorithm

In on-line handwriting, when somebody writes a word on the screen, simultaneously, the handwriting is stored as series of input points, including X and Y coordinates. As we put the pen tip down until picking it up, its movement is a continuous curve, containing series of points, which we define it as a stroke. Persian words could contain one or more stroke. For identifying of segmentation points a specific word, we need to detect all the segmentation points for each stroke [13].

Our new suggested method include 2 phase. In first phase, segmentation points are detected according to described feature. Then, in second phase, some extra points are removed by post-processes.

In first phase, following steps execute for all strokes in a word:

- A grid, including rows and columns, is drawn around the word. each small square part of grid called a Cell. Cells dimension defined by a threshold.

- If there is an input point in a cell position, that cell will be called Active cell. For all points of stroke, related cell turned to Active cell.

- Columns with only one active cell are detected. Consecutive columns with one active cell considered as integrated set. Each set, whether it is integrated set or one column, is considered as a candidate line.

- If two candidate lines are separated by one column, number of active cell on that column must be check. If number of active cell, width, is 2 then two candidate line will be joined together and will be considered as one candidate line.

- Most left column of each candidate line, consider as segmentation area. Last point of stroke that is placed in this column, will be detected as segmentation point.

By execution of these steps, all points that have described feature will be detected. To avoiding input error, in step 4, we check width of column that separated two candidate lines. If it was small, we reject it and join two candidate lines together.

In second phase, some specific extra points are removed by post-processes. Following post-processes execute, after first phase completion:

- By usage of horizontal projection concept, word baseline is detected. So by using of drawn gird, the raw that has most active cells will considered as baseline. In several researches Horizontal Projection was used to detect Persian words baseline [14].

- Considering last detected segmentation point in each stroke. If the detected point is placed below of baseline, 4 time more than threshold, while last point of stroke is placed around baseline, the detected segmentation point is considered as extra point in end of letters such as "ل" ,"ن" and etc. and will be removed.

- Considering last detected segmentation point in each stroke. If the detected point is placed around baseline, while last point of stroke is placed around same column and below of baseline, 5 time more than threshold, the detected

Table 2. Result of detected extra points by proposed method

Words	Number of words recognized without extra point	Number of words recognized with 1 extra point	Number of words recognized with 2 extra point	Number of words recognized with 3 extra point	Number of words recognized with 4 extra point
هرچند	48	2	0	0	0
تلافی	46	4	0	0	0
عاشق	42	8	0	0	0
بصیرت	49	1	0	0	0
کرج	35	13	2	0	0
مریض ترسناک	12	25	8	5	0
فلسطین	42	8	0	0	0
ملاحظه	50	0	0	0	0
گهگاه	15	17	12	5	1
تفریط	50	0	0	0	0
20 other words	834	159	7	0	0
Sum	1223	237	29	10	1

segmentation point is considered as extra point in end of the letter "م" and will be removed.

- Considering last detected segmentation point in each stroke. If the detected point is placed right side of last point of stroke, less than quadruple threshold, while end point of stroke is placed top of detected point, less than sextuple threshold, the detected segmentation point is considered as extra point in end of letters such as "ب" ,"ت" and etc. and will be removed.

- Considering last detected segmentation point in each stroke. If the detected point is placed below of baseline, 3 time more than threshold, and is one of four last stroke points, probably it is an extra point on the letter "ر". So previous points are tracked to find a point around baseline, the point is called candidate point. If there is not another segmentation point around candidate point, then detected point is considered as extra point in end of "ر" and will be removed and candidate point is considered as last segmentation point of stroke, otherwise just detected point is considered as extra point and will be removed.

- In last post-process, detected segmentation points that are placed in last or first column of a stroke, are considered as extra points and will be removed. In first phase, these points were placed in an isolate cell, so they were detected as segmentation points by mistake.

By execution of these post-processes, most of specific extra points will be removed and only correct segmentation points will be remained. Figure 2 shows detected segmentation points of the word "ملاحظه" after execution of proposed method.

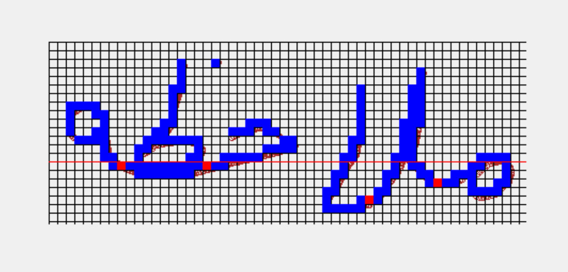

Figure. 2 Detected segmentation points of the word "ملاحظه" by proposed method

5. EXPERIMENTAL RESULTS

To evaluate the efficiency of the proposed method, we developed a program to take people handwriting by a digital pen and detect segmentation points of them by our proposed method. We used Microsoft Visual Studio 2010 programming Environment and C# programming language to develop it.

In selection of the words, we picked out 30 words that include all structures of Persian letters. We ask 50 different people to write the words by considering Standard Persian Handwriting that we described before. Table 1 presents the results of the proposed method.

Table 1. Result of detected segmentation points by proposed method

Words	Number of expected point In 50 sample	Number of detected points	Number of undetected points
هرچند	150	149	1
تلافی	150	149	1
عاشق	200	199	1
بصیرت	200	199	1
کرج	50	49	1
مریض ترسناک	400	398	2
فلسطین	350	350	0
ملاحظه	200	200	0
گهگاه	150	150	0
تفریط	150	150	0
20 other words	3150	3150	0
Sum	5150	5143	7

The experimental results show that in 5 words, only in one handwriting sample from 50 samples, one segmentation point does not detect. And also for the word "مریض ترسناک", only in two handwriting sample from 50 samples, one segmentation point does not detect. All other expected segmentation points recognized correctly by our proposed method.

Table 3. Result of detected segmentation points by our previous method

Words	Number of expected point In 50 sample	Number of detected points	Number of undetected points
هرچند	150	150	0
تلافی	150	149	1
عاشق	200	200	0
بصیرت	200	200	0
کرج	50	50	0
مریض ترسناک	400	400	0
فلسطین	350	349	1
ملاحظه	200	200	0
گهگاه	150	149	1
تفریط	150	150	0
20 other words	3150	3148	2
Sum	5150	5145	5

In addition, according to the experimental results, our proposed method could detect some extra points too. Based on the obtained result of detected extra points that is shown in Table 2, 81.53% of words recognize without any extra point and 15.8% of words recognize with one extra point and 2.67% of words recognized with 2, 3 or 4 extra points.

Also, we used handwriting samples to evaluate our previous method [13]. To improve that method, we first filtered input points and only kept adjacent points that have interspace longer than a threshold. Then, we applied our previous method [13] on filtered points. The results of our previous method for detected segmentation points and detected extra points were shown in Table 3 and Table 4.

Table 4. Result of detected extra points by our previous method

Words	Number of words recognized without extra point	Min Number of detected extra point	Max Number of detected extra point
هرچند	0	1	9
تلافی	0	1	8
عاشق	0	1	8
بصیرت	0	1	9
کرج	0	1	10
مریض ترسناک	0	1	10
فلسطین	0	1	10
ملاحظه	0	1	6
گهگاه	0	1	19
تفریط	10	1	8
20 other words	7	1	18
Sum	17		

The obtained result showed that while previous method could recognize 99.9% of segmentation points, it recognize many extra points too. Only 1.13% of words recognized without any extra points and most of words were detected with more than 3 extra points.

Table 5 shows total number of detected extra points for proposed method and our previous method. The result showed that proposed method reduced detected extra points 93.73% compared to previous [13] method.

Table 5. Total number of detected extra points by proposed method and previous method

	Total number of detected extra points
Proposed method	329
Previous [13] method	5249

6. CONCLUSIONS AND FUTURE WORKS

In this paper, due to different width in various parts of Persian letters, we proposed a new approach to detecting of segmentation points in on-line Persian cursive script words. At first, we reviewed some related works and our previous method. Then Standard Persian Handwriting was introduced. Later, specific feature of Persian language was described as well as our proposed method. The experimental results from evaluation of our algorithm have shown that 99.86% of expected segmentation points, means the last point of each letter in a joining or last point of basic shapes, are detected. Proposed method detected 81.53% of words without extra points. Comparison has indicated that proposed method reduced detected extra points 93.73% compared to our previous method.

In future works, to improve proposed method, according to detected extra points in diacritic marks of words same as slant and etc., we could separate main body of word and diacritic marks and reduced number of extra points. For future work, according to 99.86% accuracy in detection of segmentation points, we could try to recognize each detected segment by isolate Persian character recognition approaches or HMM and finally complete on-line Persian cursive word recognition.

7. REFERENCES

[1] Yuan, A., Bai, G., Yang, P., Guo, Y., & Zhao, X. 2012. Handwritten English Word Recognition based on Convolutional Neural Networks. In Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on (pp. 207-212). IEEE.

[2] Liu, C. L., Yin, F., Wang, D. H., & Wang, Q. F. 2012. Online and offline handwritten Chinese character recognition: benchmarking on new databases. Pattern Recognition.

[3] Dai, R., Liu, C., and Xiao, B. 2007. Chinese character recognition: history, status and prospects. Frontiers of Computer Science in China, 1(2), 126-136.

[4] Biadsy, F., El-Sana, J., and Habash, N. 2006. Online arabic handwriting recognition using hidden markov models. In Tenth International Workshop on Frontiers in Handwriting Recognition.

[5] Halavati, R., Jamzad, M., and Soleymani, M. 2005. A novel approach to persian online hand writing recognition. In Proceedings of the 4th World Enformatika Conference (WEC 05). Vol. 6, pp. 232-236.

[6] Pirnia, Sh., Khademi, M., Nikookar, A. and Bani, Z. 2010. A Feature-Based Approach to Segmentation of Persian Online Cursive Script. The 2010 International Conference on Computer and Software Modeling (ICCSM).

[7] Daryoush, K. S., Khademi, M., Nikookar, A., & Farahani, A. 2012. Segmentation of Persian Cursive Words Using Basic Shapes. Journal of Engineering Research and Applications (IJERA).

[8] Izadi, S., Haji, M., and Suen, C. Y. 2008. A new segmentation algorithm for online handwritten word recognition in Persian script. In Proc. Eleventh International Conf. Frontiers in Handwriting Recognition (CFHR 2008) . pp. 598-603.

[9] Harouni, M., Mohamad, D., & Rasouli, A. 2010. Deductive method for recognition of on-line handwritten Persian/Arabic characters. In Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on IEEE. Vol. 5. 791-795.

[10] Daifallah, K., Zarka, N., & Jamous, H. 2009. Recognition-based segmentation algorithm for on-line arabic handwriting. In Document Analysis and Recognition. 10th International Conference on. IEEE. 886-890.

[11] Maliki, M., Jassim, S., Al-Jawad, N., & Sellahewa, H. 2012. Arabic handwritten: pre-processing and segmentation. In Proc. of SPIE. Vol. 8406.

[12] Harouni, M., Mohamad, D., Rahim, M. S. M., and Halawani, S. M. 2012. Finding Critical Points of Handwritten Persian/Arabic Character. IJMLC.

[13]] Khademi, M., Golestani, M.R., Nikookar, A., and Farahani, A. 2013. A New Method for Detecting Segmentation Points in Persian Cursive Words. Journal of Basic and Applied Scientific Research (JBASR),Special Issue (1).

[14] Nagabhushan, P., and Alaei, A. 2010. Tracing and straightening the baseline in handwritten persian/arabic text-line: A new approach based on painting-technique. The Proceeding of Intl Journal on Computer Science and Engineering. 907-916.