PERSIAN HANDWRITTEN WORDS DETECTION BASED ON FEATURES EXTRACTION AND FUZZY ALGORITHM

Gholamreza Hadidi[1] and Hadi Delavari[2]

[1,2]Department of Electrical Engineering, Hamedan University of Technology, Hamedan, Iran

ABSTRACT

In this paper, extracted word features have been used toward offline detection of Persian handwritten words. This has been done using fourteen features in Persian word structure. In this method after required pre-processing, numbers and location of dots versus base line are extracted then the word becomes thin and the feature extraction stage starts. In this stage the words are detected by using extracted features and fuzzy algorithm .In this paper, a practical case has been investigated, in which we use a group of words including names of Iran provinces. The proposed method is faster in comparison to the common algorithms due to the fact that there is no learning procedure in the proposed method. According to the collected database, the presented method is able to detect %82.3 of handwritten Persian words.

KEYWORDS

Fuzzy algorithm, Membership function, Handwritten Recognition, Persian OCR

1.INTRODUCTION

In recent years, detection of Persian handwriting has been an important research topic [1]. The importance of this topic is attributed to the fact that the detection of handwritten words has a various and wide usage in industry. These usages include reading registration forms, sorting postal packs, reading texts in books and providing search options and categorizing texts. In general, detection of letters and words is divided in two types, online & offline detection .In online mode [2] the words have been written and detected simultaneously. But in offline mode [3], detection is completed subsequently after words are written. In the proposed method we need to extract fourteen defined features from words. Also feature extraction methods are divided into two categories, Structural methods and statistical methods. In structural methods like [3] some features related to structure of letters or words are extracted. For instance length to width ratio of letter. But in statistical methods, some other features such as number of crossovers are extracted.

In addition there are three common ways in order to perform words or letters detection, syntactic ways, structural ways and ways based on decision theory like [4] [5] [6] [7]. In detection of handwritten, decision theory is mostly used due to various shape and irregularity. This type of detection needs learning the procedure and suffers from numerous problems such as proper learning, time consumption and dependency to learning data.

DOI : 10.14810/elelij.2015.4208

Besides, in this method separating is used and usually separating lack of accuracy leads to considerable errors. The presented algorithm is used for offline detection of Persian handwriting and we used both structural & statistical methods for the purpose of detection. In addition there is no need for learning procedure. Also thinning function has been used in order to reduce the volume of data and increase speed of the process.

Reference [8] presents a new method for segmenting cursive handwritten Persian/Arabic words. The aim of [9] is to develop a fuzzy rule based expert classification system that is able to imitate human reasoning and incorporate the analyst's knowledge of seismic event classification. Reference [10] presents a new comprehensive database for isolated offline handwritten Farsi/Arabic numbers and characters for use in optical character recognition research. Reference [11] presents a cursive character recognizer, a crucial module in any Cursive Script Recognition system based on a segmentation and recognition approach. Reference [12] presents a method for Circle detection on images. In [13] A Farsi font recognition algorithm based on the fonts of some frequent text samples is proposed. Reference [14] introduces a reliable segmentation technique for Arabic handwritten script. Reference [15] explains font recognition for Persian and Arabic languages.

Section II discusses common problems in Persian handwritten letters and words detection. Section III describes the proposed algorithm correspondingly using fuzzy algorithm has been explained in section IV and simulation results has been brought in section V. Finally section VI is allocated to conclusion and detection percentage.

2. COMMON PROBLEMS IN PERSIAN HANDWRITTEN DETECTION

Every language has some difficulties and detection challenges due to rules of writing and types of alphabets. This issue becomes more important specifically in handwritten case. Persian language includes some features which makes some difficulties in detection procedure. Some of these difficulties are related to sub words which are made of connection of letters .For instance "Kerman" in Persian "کرمان" is made of three sub words "کر", "ما" & "ن". Also on of common challenges is overlapping of letters in a single word. For example two letter are connected in one side which makes it difficult for detection (as shown in Figure 1).

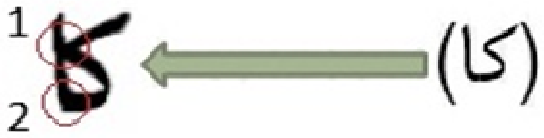

Figure 1. Irregularity in letters connection

Another example is combination of some letters. Figure 2 shows some of these combinations.

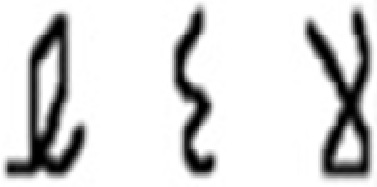

Figure 2. Letters combinations

In addition Persian handwritten has many problems which some of them are brought below:

- In Persian manuscripts, all letters are written in the fixed place above or under a straight line, which is called the baseline.
- Dots in Persian letters are very important. 56% of Persian letters have dots. Some Persian letters differ from each other just in the position and number of their dots. Some Persian letters do not have any dot. Some letters have only one dot. The others have two or three dots. These dots can be mistaken by a noise in a scanned text.
- Sometimes in handwritten and typed Persian texts, letters may have overlapped, and therefore in their separation and recognition some mistakes may happen.
- Some letter forms in Persian text have closed curves in their shapes, while others don't. This characteristic can be used as a feature to recognize some letters.
- There are different script styles for Persian language, e.g. Naskh, Nastaligh, Koofi, etc. The shapes of some letters are very different in these styles.
- Persian handwriting varies from person to person and is different from printed versions.
- There are no regular and fixed patterns to handwriting and almost each person has a different handwriting.

3. PROPOSED ALGORITHM

The proposed algorithm consists of three major parts, pre-processing, main process and detection. Initially scanned picture is applied to system as input. Then pre-processing, applies some filters and makes the picture binary. Also base line is located, numbers and locations of dots are extracted. In addition thinning occurs and skeleton of the word is extracted. In the next stage, features of pictures are extracted. These features include 14 defined features and the algorithm can detect Persian words by using these features. Table 1 describes these 14 features .The features number 4, 6 and 9 are defined firstly in this paper and have considerable contribution in detection procedure. It is true to say that defining these features and using numerical value for them are the main parts of this algorithm.

The authors collected more than 5000 handwritten sample of 31 province names in Iran. By categorizing the numerical values of these features and using fuzzy algorithm, system is able to detect handwritten Persian words.

These 14 features are defined as below:
1-Number of circles in a word
2-Area of word inscribed rectangular
3-Sum of sub words in the word
4-The maximum value of sub words inscribed rectangular
5-Sum of sub words areas
6-Area of the biggest sub word
7-Center of word location
8-The maximum value of histogram
9-Ratio of histogram maximum value to word length
10-Percentage of above pixels versus baseline
11- Percentage of below pixels versus baseline
12-Number of word parts
13-Number of pixels with 3 neighbourhood
14-Number of pixels with 4 neighbourhood

In the features above, the ones related to area are variable and dependent on resolution of the picture.

Table 2 shows the value numbers of some features.

The numerical values of these fourteen features have been extracted for all pictures of data base. Then the average value for each feature has been introduced by considering these features, the names of 31 province can be detected in an exclusive way. For example Figure 3 shows the N.P extraction in "قـز ویـن" word. This word is divided into 7 parts according to the vicinity rules and image processing functions.

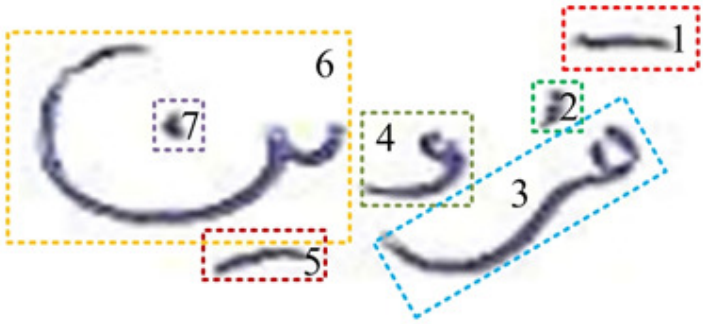

Figure 3. Word segmentation

Also Figure 4 shows the U.P, D.P and N.C extraction. In this figure base line has been detected due to the fact that number and location of dots are really important in Persian writing.
Also the number of circles equals 2 in this figure, but it depends on type of handwritten and autography which makes detection more difficult.

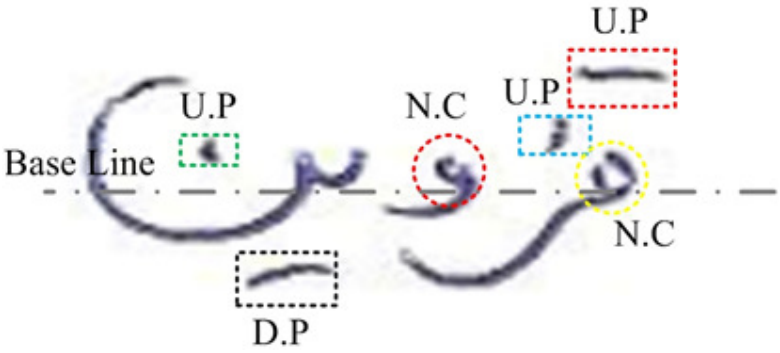

Figure 4. Base-line and some features

3.1. Pre-Processing

In this stage some initial processing is applied to the picture in order to prepare it for the main process procedure. Firstly, decision threshold function is applied then by considering the level of threshold value, the picture is converted to a binary picture. Furthermore, some filters are applied for the noise reduction purpose.

3.2. Main Process

In this stage the picture is monitored in order to extract all defined features. Then by using these features detection can be completed.

Firstly the baseline must be detected .Due to this purpose we use horizontal histogram. The baseline is located everywhere that attached to maximum number of black pixels, so the target raw is selected. Next, dots must be located.

For dots detection we should determine if the word consists of monolithic parts. In a monolithic case, we can assume that the word has no dot. But in another case, we can identify dots by comparing the length to width ratio of sub words value. Also the location of dots is compared versus base line. Then dots are deleted and word body remains. Then body is thinned in order to obtain word skeleton. For instance Figure 5 shows the thinning picture of "خراسان جنوبی".

Figure 5. Thinning output

Table 1. The extracted features

Defined Code	Feature name	Numerical range
N.C	Number of Circles	0-4
A	Area	depend on
S_Box	Sum of Boxes (inscribed rectangular)	depend on
M_Box	Maximum value of sub words inscribed rectangular	depend on
S_Box_Area	Sum of Boxes Area	depend on
M_Box_Area	Area of the biggest sub word	depend on
Cen	Center of word	5-88
Max_Proj	Maximum of Projection	4-12
Dim_ratio	Dimension Ratio	1-20
U.P	Up Pixels Percent	1-12
D.P	Down Pixels Percent	0-10
N.P	Number of Parts	3-20
NB3	Number of Pixels with 3 neighbors	2-500
NB4	Number of Pixels with 4 neighbors	0-100

The thinning function decreases the thickness of the word to just one pixel and has to maintain the overall structure of the word. For example the thinning procedure should protect teeth from destruction. This procedure really helps the system to perform with more speed due to the reduction in data volume. This thinning is performed by using presented method in [16]. This reference uses 30 masks to complete thinning function. 20 masks for boundary pixels elimination and 10 masks to retrieve the important eliminated pixels.

Following, focuses on explaining feature extraction.

3.3. Feature Extraction

The target features have been presented in Table 1 & Table 2 should be extracted. Initially close curves in word skeleton have been identified [3]. Then they must be eliminated. For instance the proposed algorithm tries to identify circles in words. Three circles have been identified which seem to be corrected but we should know in proper writing, the desired word has only one circle on "و".

This issue rises due to connection of "ج" & "خ". This usually happens in handwritten texts and also there are a lot of similar problems.

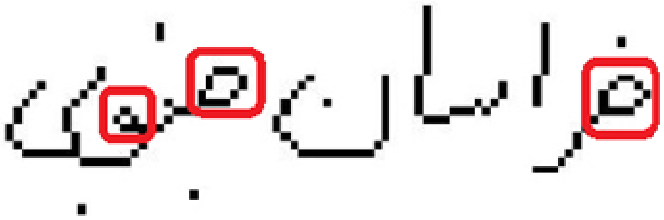

Figure 6. Circles detection

Next step is calculating the area of word inscribed rectangular. This performed in two ways in the proposed method. Figure 7 & Figure 8 show the inscribed rectangular of words & sub words respectively. After calculating the area of each sub word inscribed rectangular, the resulted values, are added together in order to form the 5th feature in table 1. Then the three novel features have been proposed in this work which two of them are associated with word area. The first is maximum value in aspect ratio of inscribed sub word rectangular which is notified with "Max_Box" in table 1.

Figure 7. Inscribed rectangular area

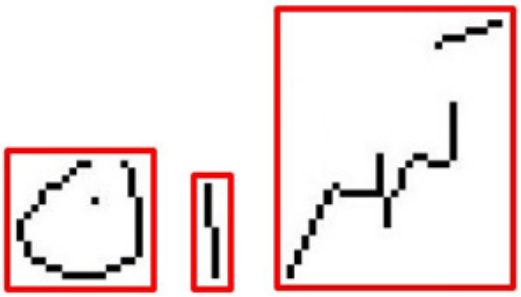

Figure 8. Inscribed rectangular area for sub-words

The Second is the area of the biggest sub word inscribed rectangular which is notified with "Max_Box_Area" in table I. Finally the third one is extracted from histogram figure & words length which is defined as ratio of histogram maximum to word length and notified with "Dim_ratio". It is worth to say that detection of base-line in pre-processing stage plays a key role in features extractions. With aim of the base-line location and by considering total pixel numbers above and below of this line, we can calculate the percentage of above and below pixels versus total pixels of the image. The following equations present these calculations.

$$\text{Pixel Counter} = \text{Up\_Counter} + \text{Down\_Counter} \quad (1)$$

$$\text{Up\_Percent} = (\text{Up\_Counter}) * 100 / \text{Pixel Counter} \quad (2)$$

$$\text{Down\_Percent} = (\text{Down\_Counter}) * 100 / \text{Pixel Counter} \quad (3)$$

In addition, number and locations of dots are significant. For instance the word "تـهـران" in Figure 7 has three dots which are all located above base-line. But dots detection is very difficult since usually people write two or three dots as a little continuous line or curve. This doubles the challenge in dots detection. For example in Figure 7, we have to decide that the little line means dots. So for more accurate dots detection we have to obtain the area of all sections (illustrated in Figure 3).

Identification of numbers and location of dots are really important in Persian handwritten detection. Regarding the fact that dots are separated from word main structure, it is better to label black pixels initially. This can be done in a way that word structure has been labelled as "1" and the other separated pixels are labelled as "2", "3", "4", … consequently. For this end we can use bwlabel syntax in Matlab. This function works in a way which identifies black pixels from the most left column and the first row. When catches the first black pixels it as "1" and eight vicinity pixels are identifiable with the same label i.e. "1". So all of the connected pixels are labelled. Then these labelled pixels can be deleted from the original image and the same procedure will be repeated to remained parts. In the next step pixels will be labelled as "2" and this procedure will be repeated till all parts are eliminated. Now by considering the numbers of pixels in each label, we allocate numbers to them appropriately. The biggest range of connected pixels are labelled "1" and other connected pixels are labelled consequently. Since the main structure pixel of the word has the maximum numbers of pixels, this part is labelled as "1", so by eliminating the label "1" pixels and saving them in a matrix, word main structure and dots are separated from each other. Label "2", "3" and … are allocated to dots and if there is a sub word with no dots, the dots matrix will be zero. Also the locations of dots are identifiable, regarding the base-line position. This can be done, using (4).

```
If min (Row) > Baseline
        Dots_Position = Up
Else                                                    (4)
        Dots_Position = Down
End
```

In this equation “baseline” illustrated the position of base-line. Dots are below the base-line if the longest row of dots matrix is smaller in comparison to base-line position and dots are above base-line if the shortest row of dots matrix is bigger than base-line position. With detection of dots positions versus base-line, the challenge rises here which will be defined as follows. In the most Persian hand written texts, the writer uses a straight line to write two dots and a curve to write three dots. According to numerous experiments, we can conclude that if the total area of each labelled part is smaller than all of total pixels, we can consider that part as possible dots. This part is a candidate for being considered as dots. Now the problem here is the fact that the candidate part is one, two or three dots. For finding the number of dots we use three factors. Number of black pixels, the covered area which dots matrix illustrated and length to width ratio of candidate part. For calculating the length to width ratio candidate part we can obtain minimum and maximum of the row and column related to candidate part. Then by differentiating them the length and width will be resulted as demonstrated in following equations.

$$\text{Length of word} = \text{Max (Column)} - \text{Min (Column)} \quad (5)$$

$$\text{Width of word} = \text{Max (Row)} - \text{Min (Row)} \quad (6)$$

$$\text{Rate} = \text{Length of word} / \text{Width of word} \quad (7)$$

Regarding the three factors above, the required rules for detection of candidate part number can be written as below:

1. If the candidate part has just one pixel or its area is smaller than 2.5, then candidate part has just one dot.
2. If length to width ratio of candidate part is bigger than 2 and the range of its area is between 2.4 to 7 then this part has two dots.
3. If candidate part area is between 7 & 10 then this part has three dots.

The following equations show the mentioned explanation.

```
If     Length (Row) == 1   ||  Area_Part < 2.5
          Number of points = 1
Elseif    Rate > 2  || Area_Part >=2.5 && area_part <=7          (8)
          Number of points = 2
Elseif   Area_Part >7 && Area_Part <10
          Number of points = 3
```

The mentioned numbers in above equations are based on experience.

4.FEATURES CATEGORIZING WITH FUZZY ALGORITHM

We have used Fuzzy algorithm in order to categorize extracted features in previous stage. After extracting the numerical values of features and by sorting these values, some words can be

detected easily using some noticeable features values. For instance the word "قم " is detectable by just one feature since this word has really short length, two circles and two parts. But in order to reach an appropriate and high detection percentage we have used all of fourteen features in Fuzzy algorithm. This guarantees efficiency for the system. For example, the Fuzzy rule for detection of word "البرز" is as it's mentioned below:

"IF (num_part is alborz) and (num_cir is alborz) and (area is alborz) and …….. and (nb4 is alb) THEN (state is alborz)"

In condition of above rule, "alborz" for each feature presents the range of numerical values for extracted features of word "البرز". For other words this procedure is repeated.

SUGENO is used as type of Fuzzy system and input membership functions are trapezoid type which range of these membership functions are proportional to minimum and maximum values of extracted features. For example the average values of some features for thirty one words are brought in Table 2.

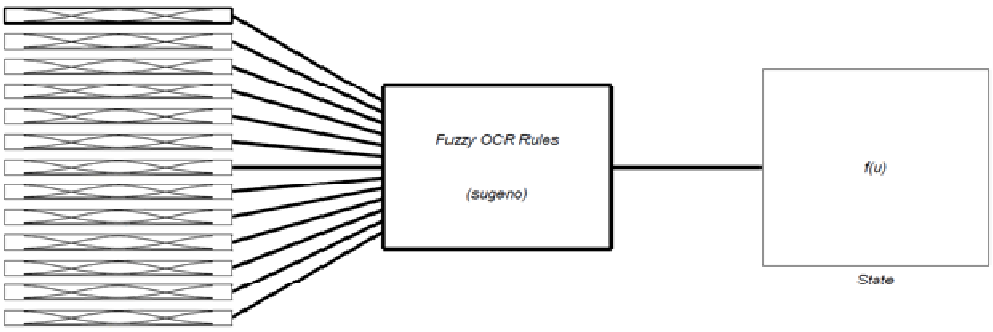

Figure 9. Fuzzy system and input variables

The inputs of Fuzzy system are fourteen extracted features and outputs are thirty one states in this case. Each state presents one province name. Figure 10 shows a sample of selected membership function which is used in this system.

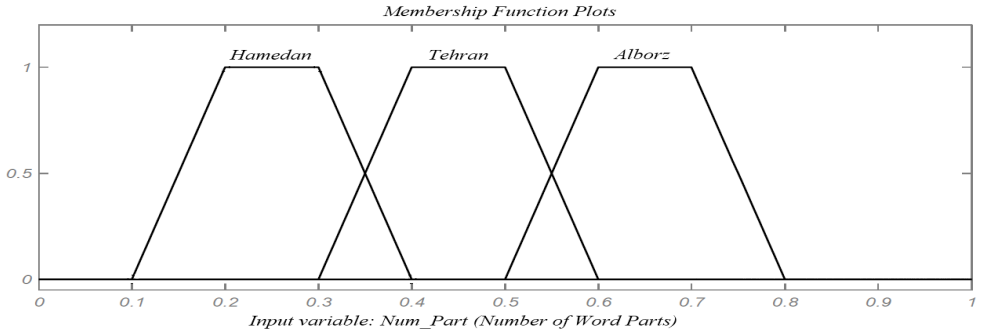

Figure 10. Sample of selected membership functions

Table 2. Average values for some features

Max_Box	Max_Box_Area	Dim_Ratio	Province name
43.06	46.42	0.2779	Hamedan
59.73	26.99	0.6045	Kermanshah
50.48	36.81	0.3815	Tehran

30.46	30.65	0.0376	Ghom
55.66	45.29	0.4658	Ardebil
35.63	22.13	0.1218	Yazd
65.94	34.37	0.5690	Mazandaran

5.SIMULATION RESULTS

MATLAB has been used in order to stimulate the proposed algorithm and this algorithm used the collected database 5000 Persian handwritten samples of Iran provinces names. The detection percentage of the proposed algorithm has achieved a percentage of 82.3 which shows appropriate accuracy with relatively high speed procedure in comparison to other methods. Table 3 shows percentage of detection with proposed algorithm for 31 words of iran provinces names.

Table 3. Percentage of detection for iran province words

Persian name of province word	English name of province word	Percentage of detection
البرز	Alborz	75.8 %
اردبیل	Ardabil	80.3 %
آذربایجان غربی	West Azerbaijan	77.6 %
آذربایجان شرقی	East Azerbaijan	77.4 %
بوشهر	Bushehr	85.1 %
چهارمحال و بختیاری	Chaharmahal and Bakhtiari	90.2 %
اصفهان	Isfahan	85.8 %
فارس	fars	97.7 %
قزوین	Qazvin	88.1 %
قم	Qom	98.9 %
گیلان	Gilan	71.8 %
گلستان	Golestan	73.9 %
همدان	Hamedan	81.3 %
هرمزگان	Hormozgan	71.1 %
ایلام	Ilam	94.3 %
کرمان	Kerman	74.9 %
کرمانشاه	Kermanshah	71.6 %
خراسان جنوبی	South Khorasan	75.2 %
خراسان رضوی	Razavi Khorasan	73.8 %
خراسان شمالی	North Khorasan	74.7 %
خوزستان	Khuzestan	76.7 %
کهگیلویه و بویر احمد	Kohgiluyeh and Boyer-Ahmad	94.6 %
کردستان	Kurdistan	79.5 %
لرستان	Lorestan	77.3 %
مرکزی	Markazi	89.4 %
مازندران	Mazandaran	86.5 %
سمنان	Semnan	79.9 %
سیستان و بلوچستان	Sistan and Baluchestan	92.8 %
تهران	Tehran	83.3 %
یزد	Yazd	95.1 %
زنجان	Zanjan	76.7 %
Average of detection percentage		82.3 %

6. CONCLUSIONS

Extracted word features have been used for offline detection of Persian handwritten words. This has been done using fourteen features in Persian words structure .In this method after required pre-processing, numbers and location of dots versus base line are extracted then the word becomes thin and the feature extraction stage starts. In this stage the word is detected by using extracted features and fuzzy algorithm. In this paper, a practical case has been investigated in which we use a group of words including names of provinces in Iran. The proposed method is faster compared with common algorithms due to the fact that there is no learning procedure in the proposed method. According to the collected database, the presented method is able to detect %82.3 of handwritten Persian words.

REFERENCES

[1] Z. Kamranian, S.A. Monadjemi, & N. Nematbakhsh, (2013) "A novel free format Persian/Arabic handwritten zip code recognition system", Computers & Electrical Engineering, Vol. 39, Issue 7, pp1970-1979.

[2] A. Malaviya, C. Leja, L. Peters, (1996) "Multi-script handwriting recognition with FOHDEL", Proceedings of NAFIPS'96, IEEE Press, Berkeley, pp147-151.

[3] Sh. Ensafi, M. Eshghi and M. Naseri, (2009) "Recognition of separate and ad joint Persian letters using primitives", Proceedings of IEEE Symposium on Industrial Electronics & Applications, Vol. 2, Kuala Lumpur, pp611-616.

[4] MR. Keyvanpour, R. Azmi, Z.S.M. Tabatabai, Z. Abdolhosseini (2014) "Handwriting Persian character recognition using optimized structural elements", Global Journal of Technology, vol. 4, No. 2, pp107-113.

[5] Y. Wang, Q. Fu, X. Ding, Ch. Liu, (2015) "Importance sampling based discriminative learning for large scale offline handwritten Chinese character recognition", Pattern Recognition, Vol. 48, Issue 4, pp1225-1234.

[6] J. Al Abodi, Xue Li , (2014) "An effective approach to offline Arabic handwriting recognition", Computers & Electrical Engineering, Vol. 40, Issue 6, pp1883–1901.

[7] A. Montaser Awal, H. Mouchère & Ch. Viard-Gaudin, (2014) "A global learning approach for an online handwritten mathematical expression recognition system", Pattern Recognition Letters, Vol. 35, pp68-77.

[8] M. Harouni, M.S.M. Rahim, M. Al-Rodhaan, T. Saba, A. Rehman, A. Al-Dhelaan, (2014) "Online Persian/Arabic script classification without contextual information", The Imaging Science Journal, vol. 62, Issue 8, pp437-448.

[9] El.H. Ait Laasri, Es-S. Akhouayri, D. Agliz, D. Zonta & A. Atmani, (2015) "A fuzzy expert system for automatic seismic signal classification", Expert Systems with Applications, Vol. 42, Issue 3, pp1013-1027.

[10] S. Mozaffari, K. Faez, F. Faradji, M. Ziaratban and S.M. Golzan, (2006) "A comprehensive isolated Farsi/Arabic character database for handwritten OCR research", Proc. 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR), La Baule, France, pp385-389.

[11] F. Camastra, A. Vinciarelli, (2003) "Combining neural gas and learning vector quantization for cursive character recognition", Neuro-computing, pp147–159.

[12] E. Cuevas, V. Osuna-Enciso and D. Oliva, (2014) "Circle detection on images based on the Clonal Selection Algorithm (CSA)", The Imaging Science Journal, Vol. 63, Issue 1, pp34–44.

[13] M. Ziaratban, F. Bagheri, (2015) "Farsi Font Recognition based on The Fonts of Text Samples Extracted by SOM", Journal of mathematics and computer science, vol. 15, Issue 1, pp40-56.

[14] Mohamed A. Ali, (2015) "An Efficient Segmentation Algorithm for Arabic Handwritten Characters Recognition System", Afro-European Conference for Industrial Advancement, Advances in Intelligent Systems and Computing, vol. 334, pp193-204.

[15] H. Luqman, Sabri A. Mahmoud, and S. Awaida, (2015) "Arabic and Farsi Font Recognition: Survey", International Journal of Pattern Recognition and Artificial Intelligence, vol. 29, Issue 1.

[16] Sh. Ensafi, M. Miremadi, M. Eshghi, M. Naseri and A. Keipour, (2009) "Recognition of Separate and Ad joint Persian Letters in Less than Three Letter Sub words Using Primitives", Proceedings of Iran 17th Electrical Engineering Conference, Tehran, pp11-13.

Authors

Gholamreza Hadidi was born in Sonqor , Iran, In 1989. He received the B.Sc. degree in Electronic Engineering from Imam Reza International University, Mashhad, Iran in 2011 and M.Sc. degree in Control Engineering from Hamedan University of Technology, Hamedan, Iran, in 2015. His research interests are intelligent Control, Image Processing, Signal Processing, Pattern Recognition and CMOS integrated circuits.

Hadi Dalavari received Ph.D. degree, M.Eng. degree and B.Eng. degree in Control Engineering in 2011, 2006 and 2004, respectively. Since 2008, he has been with Department of Electrical Engineering, Hamedan University of Technology. He is the author and co-author of about 60 publications. His research interests include nonlinear control theory and applications, fractional order control, chaos control, robotics etc.